A refined convergence analysis of pDCA$_e$ with applications to simultaneous sparse recovery and outlier detection

Tianxiang Liu* Ting Kei Pong† Akiko Takeda‡§

November 3, 2018

Abstract

We consider the problem of minimizing a difference-of-convex (DC) function, which can be written as the sum of a smooth convex function with Lipschitz gradient, a proper closed convex function and a continuous possibly nonsmooth concave function. We refine the convergence analysis in [40] for the proximal DC algorithm with extrapolation (pDCA$_e$) and show that the whole sequence generated by the algorithm is convergent *without* imposing differentiability assumptions in the concave part. Our analysis is based on a new potential function and we assume such a function is a Kurdyka-Łojasiewicz (KL) function. We also establish a relationship between our KL assumption and the one used in [40]. Finally, we demonstrate how the pDCA$_e$ can be applied to a class of simultaneous sparse recovery and outlier detection problems arising from robust compressed sensing in signal processing and least trimmed squares regression in statistics. Specifically, we show that the objectives of these problems can be written as level-bounded DC functions whose concave parts are *typically nonsmooth*. Moreover, for a large class of loss functions and regularizers, the KL exponent of the corresponding potential function are shown to be 1/2, which implies that the pDCA$_e$ is locally linearly convergent when applied to these problems. Our numerical experiments show that the pDCA$_e$ usually outperforms the proximal DC algorithm with nonmonotone linesearch [22, Appendix A] in both CPU time and solution quality for this particular application.

1 Introduction

Nonconvex optimization plays an important role in many contemporary applications such as machine learning and signal processing. In the area of machine learning, for example, nonconvex sparse learning has become a hot research topic in recent years, and a large number of papers are devoted to the study of classification/regression models with nonconvex regularizers for finding sparse solutions; see, for example, [15, 18, 42]. On the other hand, in signal processing, specifically in the area of compressed sensing, many nonconvex models have been proposed in recent years for recovering the underlying sparse/approximately sparse signals. We refer the interested readers to [9, 12, 13, 16, 33, 38, 41] and references therein for more details.

*Department of Applied Mathematics, the Hong Kong Polytechnic University, Hong Kong. Email: tianxiang.liu@polyu.edu.hk.

†Department of Applied Mathematics, the Hong Kong Polytechnic University, Hong Kong. This author's work was supported in part by Hong Kong Research Grants Council PolyU153005/17p. Email: tk.pong@polyu.edu.hk.

‡Department of Creative Informatics, Graduate School of Information Science and Technology, the University of Tokyo, Tokyo, Japan. This author's work was supported in part by JSPS KAKENHI Grant Number 15K00031. Email: takeda@mist.i.u-tokyo.ac.jp.

§RIKEN Center for Advanced Intelligence Project, 1-4-1, Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan. Email: akiko.takeda@riken.jp.

In this paper, we consider a special class of nonconvex optimization problems: the difference-of-convex optimization problems. This is a class of problems whose objective can be written as the difference of two convex functions; see the monograph [37] for a comprehensive exposition. Here, we focus on the following model,

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} F(\boldsymbol{x}) := f(\boldsymbol{x}) + P_1(\boldsymbol{x}) - P_2(\boldsymbol{x}), \tag{1}$$

where f is a smooth convex function with Lipschitz continuous gradient whose Lipschitz continuity modulus is $L > 0$, P_1 is a proper closed convex function and P_2 is a *continuous* convex function. In typical applications in sparse learning and compressed sensing, the function f is a loss function representing data fidelity, while $P = P_1 - P_2$ is a regularizer for inducing desirable structures (for example, sparsity) in the solution. Commonly used regularizers include:

- $P_1(\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|_1$ and $P_2(\boldsymbol{x}) = \lambda\sum_{i=1}^n \int_0^{|x_i|} \frac{[\min\{\theta\lambda,t\}-\lambda]_+}{(\theta-1)\lambda}dt$, where $\lambda > 0$ and $\theta > 2$. This regularizer is known as the smoothly clipped absolute deviation (SCAD) function; see [15, 18];
- $P_1(\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|_1$ and $P_2(\boldsymbol{x}) = \lambda\sum_{i=1}^n \int_0^{|x_i|} \min\{1, t/(\theta\lambda)\}dt$, where $\lambda > 0$ and $\theta > 1$. This regularizer is known as the minimax concave penalty (MCP) function; see [18, 42];
- $P_1(\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|_1$ and $P_2(\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|$, where $\lambda > 0$. This is known as the ℓ_{1-2} regularizer; see [41];
- $P_1(\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|_1$ and $P_2(\boldsymbol{x}) = \lambda\mu\sum_{i=1}^p |x_{[i]}|$, where $x_{[i]}$ denotes the ith largest element in magnitude, $\lambda > 0$, $\mu \in (0, 1]$ and $p < n$ is a positive integer. We will refer to this as the Truncated ℓ_1 regularizer; see [38];
- $P_1(\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|_1$ and $P_2(\boldsymbol{x}) = \lambda\sum_{i=1}^n [|x_i| - \theta]_+$, where $\lambda > 0$ and $\theta > 0$. This is known as the Capped ℓ_1 regularizer; see [18].

Notice that while the P_2 in the SCAD and MCP functions are smooth, the P_2 in the other three regularizers mentioned above are nonsmooth.

For difference-of-convex optimization problems such as (1), the classical method for solving them is the so-called DC algorithm (DCA), proposed in [27]. In this algorithm, in each iteration, one majorizes the concave part of the objective by its local linear approximation and solves the resulting convex optimization subproblem. For efficient implementation of this algorithm, one should construct a suitable DC decomposition so that the corresponding subproblems are easy to solve. This idea was incorporated in the so-called proximal DCA [19], which is a version of DCA that makes use of the following special DC decomposition of the objective in (1) (see [28, Eq. 16] for an earlier use of such a decomposition in solving trust region subproblems):

$$F(\boldsymbol{x}) = \left[\frac{L}{2}\|\boldsymbol{x}\|^2 + P_1(\boldsymbol{x})\right] - \left[\frac{L}{2}\|\boldsymbol{x}\|^2 - f(\boldsymbol{x}) + P_2(\boldsymbol{x})\right].$$

The major computational costs of the subproblems in the proximal DCA come from the computations of the gradient of f, the proximal mapping of P_1 and a subgradient of P_2, which are simple for commonly used loss functions and regularizers. Later, extrapolation techniques were incorporated into the proximal DCA in [40]. The resulting algorithm was called pDCA$_e$, and was shown to have much better numerical performance than the proximal DCA. Convergence analysis of the pDCA$_e$ was also presented in [40]. In particular, when F is level-bounded, it was established in [40] that any cluster point of the sequence generated by pDCA$_e$ is a stationary point of F in (1). Moreover, under an additional smoothness assumption on P_2 and by assuming that a certain potential function is a Kurdyka-Łojasiewicz (KL) function, it was further proved that the whole sequence generated by pDCA$_e$ is convergent. Local convergence rate was also discussed based on the KL exponent of the potential function. However, the analysis there heavily relies on the smoothness assumption on P_2, which does not hold for many commonly used regularizers such as the Capped ℓ_1 regularizer [18] and the Truncated ℓ_1 regularizer [38] mentioned above. More importantly, as we shall see later in Section 5, the objectives of models for simultaneous sparse recovery and outlier detection can be written as DC functions whose concave parts are *typically nonsmooth*. Thus, for these problems, the analysis in [40] cannot be applied to studying global sequential convergence nor local convergence rate of the sequence generated by pDCA$_e$.

In this paper, we refine the convergence analysis of pDCA$_e$ in [40] to cover the case when the P_2 in (1) is possibly *nonsmooth*. Our analysis is based on a potential function different from the one used in [40]. By assuming this new potential function is a KL function and F is level-bounded, we show that the whole sequence generated by pDCA$_e$ is convergent. We then study a relationship between the KL assumption used in this paper and the one used in [40]. Specifically, under a suitable smoothness assumption on P_2, we show that if the potential function used in [40] has a KL exponent of $\frac{1}{2}$, so does our new potential function. KL exponent is an important quantity that is closely related to the local convergence rate of first-order methods [2, 3, 21], and we also provide an explicit estimate of the KL exponent of the potential function used in our analysis for some commonly used F. Finally, we discuss how the pDCA$_e$ can be applied to a class of simultaneous sparse recovery and outlier detection problems in least trimmed squares regression in statistics (see [31, 32]) and robust compressed sensing in signal processing (see [10] and references therein). Specifically, we demonstrate how to explicitly rewrite the objective function as a level-bounded DC function in the form of (1), and show that the KL exponent of the corresponding potential function is $\frac{1}{2}$ for many simultaneous sparse recovery and outlier detection models: this implies local linear convergence of pDCA$_e$ when applied to these models. In our numerical experiments on this particular application, the pDCA$_e$ always outperforms the proximal DCA with nonmonotone linesearch [22] in both solution quality and CPU time.

The rest of the paper is organized as follows. In Section 2, we introduce notation and preliminary results. We also review some convergence properties of the pDCA$_e$ from [40]. In Section 3, we present our refined global convergence analysis for pDCA$_e$. Relationship between the KL assumption used in this paper and the one used in [40] is discussed in Section 4. In Section 5, we describe how our algorithm can be applied to a large class of simultaneous sparse recovery and outlier detection problems. Numerical results are presented in Section 6.

2 Notation and Preliminaries

In this paper, vectors and matrices are represented in bold lower case letters and upper case letters, respectively. We use $\mathbb{R}^n$ to denote the n-dimensional Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and the Euclidean norm $\|\cdot\|$. For a vector $\boldsymbol{x} \in \mathbb{R}^n$, we let $\|\boldsymbol{x}\|_1$ and $\|\boldsymbol{x}\|_0$ denote the ℓ_1 norm and the number of nonzero entries in $\boldsymbol{x}$ (ℓ_0 norm), respectively. For two vectors $\boldsymbol{a}$, $\boldsymbol{b} \in \mathbb{R}^n$, we let "$\circ$" denote the Hadamard (entrywise) product, i.e., $(\boldsymbol{a} \circ \boldsymbol{b})_i = a_i b_i$, $i = 1, ..., n$. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, we let $\boldsymbol{A}^\top$ denote its transpose, and we use $\lambda_{\max}(\boldsymbol{A})$ to denote the largest eigenvalue of $\boldsymbol{A}$ when $\boldsymbol{A}$ is symmetric, i.e., $\boldsymbol{A} = \boldsymbol{A}^\top$.

Next, for a nonempty closed set $\mathcal{C} \subseteq \mathbb{R}^n$, we write $\mathrm{dist}(\boldsymbol{x}, \mathcal{C}) := \inf_{\boldsymbol{y} \in \mathcal{C}} \|\boldsymbol{x} - \boldsymbol{y}\|$ and define the indicator function $\delta_{\mathcal{C}}$ as

$$\delta_{\mathcal{C}}(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \boldsymbol{x} \in \mathcal{C}, \\ \infty & \text{otherwise.} \end{cases}$$

For an extended-real-valued function $h : \mathbb{R}^n \to [-\infty, \infty]$, its domain is defined as $\mathrm{dom}\, h = \{\boldsymbol{x} \in$

$\mathbb{R}^n : h(\boldsymbol{x}) < \infty\}$. Such a function is said to be proper if it is never $-\infty$ and its domain is nonempty, and is said to be closed if it is lower semicontinuous. A proper closed function h is said to be level-bounded if $\{\boldsymbol{x} \in \mathbb{R}^n : h(\boldsymbol{x}) \le \gamma\}$ is bounded for all $\gamma \in \mathbb{R}$. Following [30, Definition 8.3], for a proper closed function $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, the (limiting) subdifferential of h at $\boldsymbol{x} \in \operatorname{dom} h$ is defined as

$$\partial h(\boldsymbol{x}) = \left\{ \boldsymbol{v} : \exists\, \boldsymbol{x}^k \xrightarrow{h} \boldsymbol{x}, \boldsymbol{v}^k \to \boldsymbol{v} \text{ with } \liminf_{\substack{\boldsymbol{y} \to \boldsymbol{x}^k \\ \boldsymbol{y} \neq \boldsymbol{x}^k}} \frac{h(\boldsymbol{y}) - h(\boldsymbol{x}^k) - \langle \boldsymbol{v}^k, \boldsymbol{y} - \boldsymbol{x}^k \rangle}{\|\boldsymbol{y} - \boldsymbol{x}^k\|} \ge 0 \ \forall\, k \right\}, \tag{2}$$

where $\boldsymbol{x}^k \xrightarrow{h} \boldsymbol{x}$ means $\boldsymbol{x}^k \to \boldsymbol{x}$ and $h(\boldsymbol{x}^k) \to h(\boldsymbol{x})$. We also write $\operatorname{dom} \partial h := \{\boldsymbol{x} \in \mathbb{R}^n : \partial h(\boldsymbol{x}) \neq \emptyset\}$. It is known that if h is continuously differentiable, the subdifferential (2) reduces to the gradient of h denoted by ∇h; see, for example, [30, Exercise 8.8(b)]. When h is convex, the above subdifferential reduces to the classical subdifferential in convex analysis, see, for example, [30, Proposition 8.12]. Let h^* denote the convex conjugate of a proper closed convex function h, i.e.,

$$h^*(\boldsymbol{u}) = \sup_{\boldsymbol{x} \in \mathbb{R}^n} \{\langle \boldsymbol{u}, \boldsymbol{x} \rangle - h(\boldsymbol{x})\}.$$

Then h^* is proper closed convex and the Young's inequality holds, relating h, h^* and their subgradients: for any $\boldsymbol{x}$ and $\boldsymbol{y}$, it holds that

$$h(\boldsymbol{x}) + h^*(\boldsymbol{y}) \ge \langle \boldsymbol{x}, \boldsymbol{y} \rangle,$$

and the equality holds if and only if $\boldsymbol{y} \in \partial h(\boldsymbol{x})$. Moreover, for any $\boldsymbol{x}$ and $\boldsymbol{y}$, one has $\boldsymbol{y} \in \partial h(\boldsymbol{x})$ if and only if $\boldsymbol{x} \in \partial h^*(\boldsymbol{y})$.

We next recall the Kurdyka-Łojasiewicz (KL) property, which is satisfied by many functions such as proper closed semialgebraic functions, and is important for analyzing global sequential convergence and local convergence rate of first-order methods; see, for example, [2, 3, 4]. For notational simplicity, for any $a \in (0, \infty]$, we let Ξ_a denote the set of all concave continuous functions $\varphi : [0, a) \to [0, \infty)$ that are continuously differentiable on $(0, a)$ with positive derivatives and satisfy $\varphi(0) = 0$.

Definition 2.1. (KL property and KL exponent) *A proper closed function h is said to satisfy the KL property at $\bar{\boldsymbol{x}} \in \operatorname{dom} \partial h$ if there exist $a \in (0, \infty]$, $\varphi \in \Xi_a$ and a neighborhood U of $\bar{\boldsymbol{x}}$ such that*

$$\varphi'(h(\boldsymbol{x}) - h(\bar{\boldsymbol{x}}))\operatorname{dist}(\boldsymbol{0}, \partial h(\boldsymbol{x})) \ge 1 \tag{3}$$

whenever $\boldsymbol{x} \in U$ and $h(\bar{\boldsymbol{x}}) < h(\boldsymbol{x}) < h(\bar{\boldsymbol{x}}) + a$. If h satisfies the KL property at $\bar{\boldsymbol{x}} \in \operatorname{dom} \partial h$ and the φ in (3) can be chosen as $\varphi(s) = cs^{1-\alpha}$ for some $\alpha \in [0, 1)$ and $c > 0$, then we say that h satisfies the KL property at $\bar{\boldsymbol{x}}$ with exponent α. We say that h is a KL function if h satisfies the KL property at all points in $\operatorname{dom} \partial h$, and say that h is a KL function with exponent $\alpha \in [0, 1)$ if h satisfies the KL property with exponent α at all points in $\operatorname{dom} \partial h$.

The following lemma was proved in [6], which concerns the uniformized KL property. This property is useful for establishing convergence of first-order methods for level-bounded functions.

Lemma 2.1. (Uniformized KL property) *Suppose that h is a proper closed function and let Γ be a compact set. If h is a constant on Γ and satisfies the KL property at each point of Γ, then there exist $\epsilon, a > 0$ and $\varphi \in \Xi_a$ such that*

$$\varphi'(h(\boldsymbol{x}) - h(\hat{\boldsymbol{x}})) \cdot \operatorname{dist}(\boldsymbol{0}, \partial h(\boldsymbol{x})) \ge 1 \tag{4}$$

for any $\hat{\boldsymbol{x}} \in \Gamma$ and any $\boldsymbol{x}$ satisfying $\operatorname{dist}(\boldsymbol{x}, \Gamma) < \epsilon$ and $h(\hat{\boldsymbol{x}}) < h(\boldsymbol{x}) < h(\hat{\boldsymbol{x}}) + a$.

Algorithm 1 Proximal difference-of-convex algorithm with extrapolation (pDCA$_e$) for (1)

Input: $\boldsymbol{x}^0 \in \operatorname{dom} P_1$, $\{\beta_k\} \subseteq [0,1)$ with $\sup_k \beta_k < 1$. Set $\boldsymbol{x}^{-1} = \boldsymbol{x}^0$.

for $k = 0, 1, 2, \ldots$

take any $\boldsymbol{\xi}^{k+1} \in \partial P_2(\boldsymbol{x}^k)$ and set

$$\begin{cases} \quad \boldsymbol{u}^k = \boldsymbol{x}^k + \beta_k(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}), \\ \boldsymbol{x}^{k+1} = \underset{\boldsymbol{x}\in\mathbb{R}^n}{\arg\min} \left\{ \langle \nabla f(\boldsymbol{u}^k) - \boldsymbol{\xi}^{k+1}, \boldsymbol{x}\rangle + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{u}^k\|^2 + P_1(\boldsymbol{x}) \right\}. \end{cases} \tag{5}$$

end for

Before ending this section, we review some known results concerning the pDCA$_e$ proposed in [40] for solving (1). The algorithm is presented as Algorithm 1. This algorithm was shown to be convergent under suitable assumptions in [40]. The convergence analysis there was based on the following potential function; see [40, Eq. (4.10)]:

$$\hat{E}(\boldsymbol{x}, \boldsymbol{w}) = F(\boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{w}\|^2. \tag{6}$$

This potential function has been used for analyzing convergence of variants of the proximal gradient algorithm with extrapolations; see, for example, [11, 39]. By showing that the potential function $\hat{E}$ is nonincreasing along the sequence $\{(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k)\}$ generated by the pDCA$_e$, the following subsequential convergence result was established in [40, Theorem 4.1]; recall that $\bar{\boldsymbol{x}}$ is a stationary point of F in (1) if

$$\mathbf{0} \in \nabla f(\bar{\boldsymbol{x}}) + \partial P_1(\bar{\boldsymbol{x}}) - \partial P_2(\bar{\boldsymbol{x}}).$$

Theorem 2.1. *Suppose that F in* (1) *is level-bounded. Let $\{\boldsymbol{x}^k\}$ be the sequence generated by* pDCA$_e$ *for solving* (1)*. Then the following statements hold.*

(i) $\lim_{k\to\infty} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| = 0$.

(ii) *The sequence $\{\boldsymbol{x}^k\}$ is bounded and any accumulation point of $\{\boldsymbol{x}^k\}$ is a stationary point of F.*

Global sequential convergence of the whole sequence generated by the pDCA$_e$ was established in [40, Theorem 4.2] by assuming in addition that $\hat{E}$ is a KL function and that P_2 satisfies a certain *smoothness* condition.

Theorem 2.2. *Suppose that F in* (1) *is level-bounded and $\hat{E}$ in* (6) *is a KL function. Suppose in addition that P_2 is continuously differentiable on an open set $\mathcal{N}$ containing the set of stationary points of F, with ∇P_2 being locally Lipschitz on $\mathcal{N}$. Let $\{\boldsymbol{x}^k\}$ be the sequence generated by* pDCA$_e$ *for solving* (1)*. Then $\{\boldsymbol{x}^k\}$ converges to a stationary point of F.*

In addition, under the assumptions of Theorem 2.2 and by assuming further that $\hat{E}$ is a KL function with exponent $\alpha \in [0,1)$, local convergence rate of the sequence generated by pDCA$_e$ can be characterized by α; see [40, Theorem 4.3].

The results on global sequential convergence and local convergence rate from [40] mentioned above were derived based on the smoothness assumption on P_2. However, as we pointed out in the introduction, the P_2 in some important regularizers used in practice are nonsmooth. Moreover, as we shall see in Section 5.1, the concave parts in the DC decompositions of many models for simultaneous sparse recovery and outlier detection are typically nonsmooth. Thus, for these problems, global sequential convergence and local convergence rate of the pDCA$_e$ cannot be deduced from [40]. In the next section, we refine the convergence analysis in [40] and establish global convergence of the sequence generated by pDCA$_e$ *without* requiring P_2 to be smooth.

3 Convergence analysis

In this section, we present our global convergence results for pDCA$_e$. Unlike the analysis in [40], we do not require P_2 to be smooth. The key departure of our analysis from that in [40] is that, instead of using the function $\hat{E}$ in (6), we make use of the following auxiliary function and its KL property extensively in our analysis:

$$E(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w}) = f(\boldsymbol{x}) + P_1(\boldsymbol{x}) - \langle \boldsymbol{x},\boldsymbol{y}\rangle + P_2^*(\boldsymbol{y}) + \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{w}\|^2. \tag{7}$$

It is easy to see from Young's inequality that

$$E(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w}) \geq f(\boldsymbol{x}) + P_1(\boldsymbol{x}) - P_2(\boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{w}\|^2 = \hat{E}(\boldsymbol{x},\boldsymbol{w}) \geq F(\boldsymbol{x}). \tag{8}$$

Hence, the function E is a majorant of both $\hat{E}$ and F. Similar to the development in [40, Section 4.1], we first present some useful properties of E along the sequences $\{\boldsymbol{x}^k\}$ and $\{\boldsymbol{\xi}^k\}$ generated by pDCA$_e$ in the next proposition.

Proposition 3.1. *Suppose that F in* (1) *is level-bounded and let E be defined in* (7)*. Let $\{\boldsymbol{x}^k\}$ and $\{\boldsymbol{\xi}^k\}$ be the sequences generated by* pDCA$_e$ *for solving* (1)*. Then the following statements hold.*

(i) *For any $k \geq 1$,*

$$E(\boldsymbol{x}^{k+1},\boldsymbol{\xi}^{k+1},\boldsymbol{x}^k) \leq E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1}) - \frac{L}{2}(1-\beta_k^2)\|\boldsymbol{x}^k-\boldsymbol{x}^{k-1}\|^2. \tag{9}$$

(ii) *The sequences $\{\boldsymbol{x}^k\}$ and $\{\boldsymbol{\xi}^k\}$ are bounded. Hence, the set of accumulation points of the sequence $\{(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})\}$, denoted by Υ, is a nonempty compact set.*

(iii) *The limit $\zeta := \lim_{k\to\infty} E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})$ exists and $E \equiv \zeta$ on Υ.*

(iv) *There exists $D > 0$ such that for any $k \geq 1$, we have*

$$\mathrm{dist}(\boldsymbol{0},\partial E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})) \leq D(\|\boldsymbol{x}^k-\boldsymbol{x}^{k-1}\| + \|\boldsymbol{x}^{k-1}-\boldsymbol{x}^{k-2}\|). \tag{10}$$

Proof. We first prove (i). Using the definition of $\boldsymbol{x}^{k+1}$ in (5) as a global minimizer of a strongly convex function, we have

$$\begin{aligned}&\langle \nabla f(\boldsymbol{u}^k) - \boldsymbol{\xi}^{k+1}, \boldsymbol{x}^{k+1}\rangle + \frac{L}{2}\|\boldsymbol{x}^{k+1}-\boldsymbol{u}^k\|^2 + P_1(\boldsymbol{x}^{k+1})\\ &\leq \langle \nabla f(\boldsymbol{u}^k) - \boldsymbol{\xi}^{k+1}, \boldsymbol{x}^k\rangle + \frac{L}{2}\|\boldsymbol{x}^k-\boldsymbol{u}^k\|^2 + P_1(\boldsymbol{x}^k) - \frac{L}{2}\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^k\|^2.\end{aligned}$$

Rearranging terms in the above inequality, we see that

$$\begin{aligned}P_1(\boldsymbol{x}^{k+1}) \leq &\langle \nabla f(\boldsymbol{u}^k) - \boldsymbol{\xi}^{k+1}, \boldsymbol{x}^k - \boldsymbol{x}^{k+1}\rangle + \frac{L}{2}\|\boldsymbol{x}^k-\boldsymbol{u}^k\|^2 + P_1(\boldsymbol{x}^k)\\ &- \frac{L}{2}\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^k\|^2 - \frac{L}{2}\|\boldsymbol{x}^{k+1}-\boldsymbol{u}^k\|^2.\end{aligned}$$

Using this inequality, we obtain

$$
\begin{aligned}
&E(\boldsymbol{x}^{k+1},\boldsymbol{\xi}^{k+1},\boldsymbol{x}^k) = f(\boldsymbol{x}^{k+1}) + P_1(\boldsymbol{x}^{k+1}) - \langle \boldsymbol{x}^{k+1},\boldsymbol{\xi}^{k+1}\rangle + P_2^*(\boldsymbol{\xi}^{k+1}) + \frac{L}{2}\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^k\|^2\\
&\le f(\boldsymbol{x}^{k+1}) + \langle \nabla f(\boldsymbol{u}^k) - \boldsymbol{\xi}^{k+1}, \boldsymbol{x}^k - \boldsymbol{x}^{k+1}\rangle + \frac{L}{2}\|\boldsymbol{x}^k-\boldsymbol{u}^k\|^2 + P_1(\boldsymbol{x}^k) - \frac{L}{2}\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^k\|^2\\
&\quad - \frac{L}{2}\|\boldsymbol{x}^{k+1}-\boldsymbol{u}^k\|^2 - \langle \boldsymbol{x}^{k+1},\boldsymbol{\xi}^{k+1}\rangle + P_2^*(\boldsymbol{\xi}^{k+1}) + \frac{L}{2}\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^k\|^2\\
&= f(\boldsymbol{x}^{k+1}) + \langle \nabla f(\boldsymbol{u}^k) - \boldsymbol{\xi}^{k+1}, \boldsymbol{x}^k - \boldsymbol{x}^{k+1}\rangle + \frac{L}{2}\|\boldsymbol{x}^k-\boldsymbol{u}^k\|^2 + P_1(\boldsymbol{x}^k) - \frac{L}{2}\|\boldsymbol{x}^{k+1}-\boldsymbol{u}^k\|^2\\
&\quad + \langle \boldsymbol{x}^k - \boldsymbol{x}^{k+1}, \boldsymbol{\xi}^{k+1}\rangle - P_2(\boldsymbol{x}^k),
\end{aligned} \tag{11}
$$

where the second equality follows from $\boldsymbol{\xi}^{k+1} \in \partial P_2(\boldsymbol{x}^k)$. Now, using the Lipschitz continuity of ∇f, we see that

$$
f(\boldsymbol{x}^{k+1}) \le f(\boldsymbol{u}^k) + \langle \nabla f(\boldsymbol{u}^k), \boldsymbol{x}^{k+1} - \boldsymbol{u}^k\rangle + \frac{L}{2}\|\boldsymbol{x}^{k+1}-\boldsymbol{u}^k\|^2.
$$

Combining this with (11), we deduce further that for $k \ge 1$,

$$
\begin{aligned}
&E(\boldsymbol{x}^{k+1},\boldsymbol{\xi}^{k+1},\boldsymbol{x}^k)\\
&\le f(\boldsymbol{u}^k) + \langle \nabla f(\boldsymbol{u}^k), \boldsymbol{x}^{k+1} - \boldsymbol{u}^k\rangle + \frac{L}{2}\|\boldsymbol{x}^{k+1}-\boldsymbol{u}^k\|^2 + \langle \nabla f(\boldsymbol{u}^k) - \boldsymbol{\xi}^{k+1}, \boldsymbol{x}^k - \boldsymbol{x}^{k+1}\rangle\\
&\quad + \frac{L}{2}\|\boldsymbol{x}^k-\boldsymbol{u}^k\|^2 + P_1(\boldsymbol{x}^k) - \frac{L}{2}\|\boldsymbol{x}^{k+1}-\boldsymbol{u}^k\|^2 + \langle \boldsymbol{x}^k - \boldsymbol{x}^{k+1}, \boldsymbol{\xi}^{k+1}\rangle - P_2(\boldsymbol{x}^k)\\
&= f(\boldsymbol{u}^k) + \langle \nabla f(\boldsymbol{u}^k), \boldsymbol{x}^k - \boldsymbol{u}^k\rangle + \frac{L}{2}\|\boldsymbol{x}^k-\boldsymbol{u}^k\|^2 + P_1(\boldsymbol{x}^k) - P_2(\boldsymbol{x}^k)\\
&\le f(\boldsymbol{x}^k) + \frac{L}{2}\|\boldsymbol{x}^k-\boldsymbol{u}^k\|^2 + P_1(\boldsymbol{x}^k) - \langle \boldsymbol{x}^k,\boldsymbol{\xi}^k\rangle + P_2^*(\boldsymbol{\xi}^k)\\
&= f(\boldsymbol{x}^k) + P_1(\boldsymbol{x}^k) - \langle \boldsymbol{x}^k,\boldsymbol{\xi}^k\rangle + P_2^*(\boldsymbol{\xi}^k) + \frac{L}{2}\beta_k^2\|\boldsymbol{x}^k-\boldsymbol{x}^{k-1}\|^2\\
&= E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1}) - \frac{L}{2}(1-\beta_k^2)\|\boldsymbol{x}^k-\boldsymbol{x}^{k-1}\|^2,
\end{aligned}
$$

where the second inequality follows from the convexity of f and the Young's inequality applied to P_2, and the second equality follows from the definition of $\boldsymbol{u}^k$ in (5). This proves (i).

For statement (ii), we first note from Theorem 2.1(ii) that $\{\boldsymbol{x}^k\}$ is bounded. The boundedness of $\{\boldsymbol{\xi}^k\}$ then follows immediately from this, the continuity of P_2 and the fact that $\boldsymbol{\xi}^k \in \partial P_2(\boldsymbol{x}^{k-1})$ for $k \ge 1$. This proves (ii).

Now we prove (iii). First, we see from $\sup_k \beta_k < 1$ and (9) that the sequence $\{E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})\}$ is nonincreasing. On the other hand, since F is proper closed and level-bounded, we see that F is bounded below; see [30, Theorem 1.9]. This together with (8) implies that the sequence $\{E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})\}$ is also bounded below. Thus, $\zeta := \lim_{k\to\infty} E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})$ exists.

We next show that $E \equiv \zeta$ on Υ. To this end, take any $(\hat{\boldsymbol{x}},\hat{\boldsymbol{\xi}},\hat{\boldsymbol{w}}) \in \Upsilon$. Then there exists a convergent subsequence $\{(\boldsymbol{x}^{k_i},\boldsymbol{\xi}^{k_i},\boldsymbol{x}^{k_i-1})\}$ such that $\lim_{i\to\infty}(\boldsymbol{x}^{k_i},\boldsymbol{\xi}^{k_i},\boldsymbol{x}^{k_i-1}) = (\hat{\boldsymbol{x}},\hat{\boldsymbol{\xi}},\hat{\boldsymbol{w}})$. Now, using the definition of $\boldsymbol{x}^{k_i}$ as the minimizer of the $\boldsymbol{x}$-subproblem in (5), we have

$$
\begin{aligned}
&\langle \nabla f(\boldsymbol{u}^{k_i-1}) - \boldsymbol{\xi}^{k_i}, \boldsymbol{x}^{k_i}\rangle + \frac{L}{2}\|\boldsymbol{x}^{k_i}-\boldsymbol{u}^{k_i-1}\|^2 + P_1(\boldsymbol{x}^{k_i})\\
\le\ &\langle \nabla f(\boldsymbol{u}^{k_i-1}) - \boldsymbol{\xi}^{k_i}, \hat{\boldsymbol{x}}\rangle + \frac{L}{2}\|\hat{\boldsymbol{x}}-\boldsymbol{u}^{k_i-1}\|^2 + P_1(\hat{\boldsymbol{x}}).
\end{aligned}
$$

Rearranging terms in the above inequality, we obtain further that

$$
\langle \nabla f(\boldsymbol{u}^{k_i-1}) - \boldsymbol{\xi}^{k_i}, \boldsymbol{x}^{k_i} - \hat{\boldsymbol{x}}\rangle + \frac{L}{2}\|\boldsymbol{x}^{k_i}-\boldsymbol{u}^{k_i-1}\|^2 + P_1(\boldsymbol{x}^{k_i}) \le \frac{L}{2}\|\hat{\boldsymbol{x}}-\boldsymbol{u}^{k_i-1}\|^2 + P_1(\hat{\boldsymbol{x}}). \tag{12}
$$

On the other hand, using the triangle inequality and the definition of $\boldsymbol{u}^k$ in (5), we see that

$$\|\hat{\boldsymbol{x}} - \boldsymbol{u}^{k_i-1}\| = \|\hat{\boldsymbol{x}} - \boldsymbol{x}^{k_i} + \boldsymbol{x}^{k_i} - \boldsymbol{u}^{k_i-1}\| \leq \|\hat{\boldsymbol{x}} - \boldsymbol{x}^{k_i}\| + \|\boldsymbol{x}^{k_i} - \boldsymbol{u}^{k_i-1}\|,$$
$$\|\boldsymbol{x}^{k_i} - \boldsymbol{u}^{k_i-1}\| = \|\boldsymbol{x}^{k_i} - \boldsymbol{x}^{k_i-1} - \beta_{k_i-1}(\boldsymbol{x}^{k_i-1} - \boldsymbol{x}^{k_i-2})\| \leq \|\boldsymbol{x}^{k_i} - \boldsymbol{x}^{k_i-1}\| + \|\boldsymbol{x}^{k_i-1} - \boldsymbol{x}^{k_i-2}\|.$$

These together with $\lim_{k\to\infty} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| = 0$ from Theorem 2.1(i) imply

$$\|\hat{\boldsymbol{x}} - \boldsymbol{u}^{k_i-1}\| \to 0 \quad \text{and} \quad \|\boldsymbol{x}^{k_i} - \boldsymbol{u}^{k_i-1}\| \to 0. \tag{13}$$

In addition, using the continuous differentiability of f and the boundedness of $\{\boldsymbol{\xi}^{k_i}\}$ and $\{\boldsymbol{u}^{k_i}\}$, we have

$$\lim_{i\to\infty} \langle \nabla f(\boldsymbol{u}^{k_i-1}) - \boldsymbol{\xi}^{k_i}, \boldsymbol{x}^{k_i} - \hat{\boldsymbol{x}}\rangle = 0. \tag{14}$$

Thus, we have

$$\begin{aligned}
\zeta &= \lim_{i\to\infty} E(\boldsymbol{x}^{k_i}, \boldsymbol{\xi}^{k_i}, \boldsymbol{x}^{k_i-1}) = \lim_{i\to\infty} f(\boldsymbol{x}^{k_i}) + P_1(\boldsymbol{x}^{k_i}) - \langle \boldsymbol{x}^{k_i}, \boldsymbol{\xi}^{k_i}\rangle + P_2^*(\boldsymbol{\xi}^{k_i}) + \frac{L}{2}\|\boldsymbol{x}^{k_i} - \boldsymbol{x}^{k_i-1}\|^2 \\
&= \lim_{i\to\infty} f(\boldsymbol{x}^{k_i}) + P_1(\boldsymbol{x}^{k_i}) - \langle \boldsymbol{x}^{k_i}, \boldsymbol{\xi}^{k_i}\rangle + P_2^*(\boldsymbol{\xi}^{k_i}) + \frac{L}{2}\|\boldsymbol{x}^{k_i} - \boldsymbol{u}^{k_i-1}\|^2 \\
&= \lim_{i\to\infty} f(\boldsymbol{x}^{k_i}) + \langle \nabla f(\boldsymbol{u}^{k_i-1}) - \boldsymbol{\xi}^{k_i}, \boldsymbol{x}^{k_i} - \hat{\boldsymbol{x}}\rangle + \frac{L}{2}\|\boldsymbol{x}^{k_i} - \boldsymbol{u}^{k_i-1}\|^2 + P_1(\boldsymbol{x}^{k_i}) - \langle \boldsymbol{x}^{k_i}, \boldsymbol{\xi}^{k_i}\rangle + P_2^*(\boldsymbol{\xi}^{k_i}) \\
&\leq \limsup_{i\to\infty} f(\boldsymbol{x}^{k_i}) + \frac{L}{2}\|\hat{\boldsymbol{x}} - \boldsymbol{u}^{k_i-1}\|^2 + P_1(\hat{\boldsymbol{x}}) - \langle \boldsymbol{x}^{k_i}, \boldsymbol{\xi}^{k_i}\rangle + P_2^*(\boldsymbol{\xi}^{k_i}) \\
&= \limsup_{i\to\infty} f(\boldsymbol{x}^{k_i}) + P_1(\hat{\boldsymbol{x}}) - \langle \boldsymbol{x}^{k_i} - \boldsymbol{x}^{k_i-1}, \boldsymbol{\xi}^{k_i}\rangle - P_2(\boldsymbol{x}^{k_i-1}) \\
&= f(\hat{\boldsymbol{x}}) + P_1(\hat{\boldsymbol{x}}) - P_2(\hat{\boldsymbol{x}}) = F(\hat{\boldsymbol{x}}) \leq E(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{w}}),
\end{aligned}$$

where the third equality follows from (13) and $\|\boldsymbol{x}^{k_i} - \boldsymbol{x}^{k_i-1}\| \to 0$, the fourth equality follows from (14), the first inequality follows from (12), the fifth equality follows from $\|\hat{\boldsymbol{x}} - \boldsymbol{u}^{k_i-1}\| \to 0$ (see (13)) and $\boldsymbol{\xi}^{k_i} \in \partial P_2(\boldsymbol{x}^{k_i-1})$, the sixth equality follows from $\|\boldsymbol{x}^{k_i} - \boldsymbol{x}^{k_i-1}\| \to 0$, the boundedness of $\{\boldsymbol{\xi}^k\}$ and the continuity of f and P_2, and the last inequality follows from (8). Finally, since E is lower semicontinuous, we also have

$$E(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{w}}) \leq \liminf_{i\to\infty} E(\boldsymbol{x}^{k_i}, \boldsymbol{\xi}^{k_i}, \boldsymbol{x}^{k_i-1}) = \zeta.$$

Consequently, $E(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{w}}) = \zeta$. We then conclude that $E \equiv \zeta$ from the arbitrariness of $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{w}})$ on Υ. This proves (iii).

Finally, we prove (iv). Note that the subdifferential of the function E at the point $(\boldsymbol{x}^k, \boldsymbol{\xi}^k, \boldsymbol{x}^{k-1})$, $k \geq 1$, is:

$$\partial E(\boldsymbol{x}^k, \boldsymbol{\xi}^k, \boldsymbol{x}^{k-1}) = \begin{bmatrix} \nabla f(\boldsymbol{x}^k) + \partial P_1(\boldsymbol{x}^k) - \boldsymbol{\xi}^k + L(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}) \\ -\boldsymbol{x}^k + \partial P_2^*(\boldsymbol{\xi}^k) \\ -L(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}) \end{bmatrix}.$$

On the other hand, one can see from pDCA$_e$ that $\boldsymbol{x}^{k-1} \in \partial P_2^*(\boldsymbol{\xi}^k)$ for $k \geq 1$. Moreover, we know from the $\boldsymbol{x}$-update in (5) that for $k \geq 1$,

$$-\nabla f(\boldsymbol{u}^{k-1}) + \boldsymbol{\xi}^k - L(\boldsymbol{x}^k - \boldsymbol{u}^{k-1}) \in \partial P_1(\boldsymbol{x}^k).$$

Using these relations, we see further that for $k \geq 1$,

$$\begin{bmatrix} \nabla f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{u}^{k-1}) - L(\boldsymbol{x}^{k-1} - \boldsymbol{u}^{k-1}) \\ \boldsymbol{x}^{k-1} - \boldsymbol{x}^k \\ -L(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}) \end{bmatrix} \in \partial E(\boldsymbol{x}^k, \boldsymbol{\xi}^k, \boldsymbol{x}^{k-1}).$$

This together with the definition of $\boldsymbol{u}^k$ and the Lipschitz continuity of ∇f implies that (10) holds for some $D > 0$. This completes the proof. □

Equipped with the properties of E established in Proposition 3.1, we are now ready to present our global convergence analysis. We will show that the sequence $\{\boldsymbol{x}^k\}$ generated by pDCA$_e$ is convergent to a stationary point of F in (1) under the additional assumption that the function E defined in (7) is a KL function. Unlike [40], we do not impose smoothness assumptions on P_2. The line of arguments we use in the proof are standard among convergence analysis based on KL property; see, for example, [2, 3, 4]. We include the proof for completeness.

Theorem 3.1. *Suppose that F in* (1) *is level-bounded and the E defined in* (7) *is a KL function. Let $\{\boldsymbol{x}^k\}$ be the sequence generated by* pDCA$_e$ *for solving* (1). *Then the sequence $\{\boldsymbol{x}^k\}$ is convergent to a stationary point of F. Moreover, $\sum_{k=1}^{\infty}\|\boldsymbol{x}^k-\boldsymbol{x}^{k-1}\|<\infty$.*

Proof. In view of Theorem 2.1(ii), it suffices to prove that $\{\boldsymbol{x}^k\}$ is convergent and $\sum_{k=1}^{\infty}\|\boldsymbol{x}^k-\boldsymbol{x}^{k-1}\|<\infty$. To this end, we first recall from Proposition 3.1(iii) and (9) that the sequence $\{E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})\}$ is nonincreasing (Recall that $\sup_k\beta_k<1$) and $\zeta=\lim_{k\to\infty}E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})$ exists. Thus, if there exists some $N>0$ such that $E(\boldsymbol{x}^N,\boldsymbol{\xi}^N,\boldsymbol{x}^{N-1})=\zeta$, then it must hold that $E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})=\zeta$ for all $k\geq N$. Therefore, we know from (9) that $\boldsymbol{x}^k=\boldsymbol{x}^N$ for any $k\geq N$, implying that $\{\boldsymbol{x}^k\}$ converges finitely.

We next consider the case that $E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})>\zeta$ for all k. Recall from Proposition 3.1(ii) that Υ is the (compact) set of accumulation points of $\{(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})\}$. Since E satisfies the KL property at each point in the compact set $\Upsilon\subseteq\operatorname{dom}\partial E$ and $E\equiv\zeta$ on Υ, by Lemma 2.1, there exist an $\epsilon>0$ and a continuous concave function $\varphi\in\Xi_a$ with $a>0$ such that

$$\varphi'(E(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w})-\zeta)\cdot\operatorname{dist}(\mathbf{0},\partial E(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w}))\geq 1 \tag{15}$$

for all $(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w})\in U$, where

$$U=\{(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w}):\operatorname{dist}((\boldsymbol{x},\boldsymbol{y},\boldsymbol{w}),\Upsilon)<\epsilon\}\cap\{(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w}):\zeta<E(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w})<\zeta+a\}.$$

Since Υ is the set of accumulation points of the bounded sequence $\{(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})\}$, we have

$$\lim_{k\to\infty}\operatorname{dist}((\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1}),\Upsilon)=0.$$

Hence, there exists $N_1>0$ such that $\operatorname{dist}((\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1}),\Upsilon)<\epsilon$ for any $k\geq N_1$. In addition, since the sequence $\{E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})\}$ converges to ζ by Proposition 3.1(iii), there exists $N_2>0$ such that $\zeta<E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})<\zeta+a$ for any $k\geq N_2$. Let $\bar N=\max\{N_1,N_2\}$. Then the sequence $\{(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})\}_{k\geq\bar N}$ belongs to U and we deduce from (15) that

$$\varphi'(E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})-\zeta)\cdot\operatorname{dist}(\mathbf{0},\partial E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1}))\geq 1,\quad\forall\, k\geq\bar N. \tag{16}$$

Using the concavity of φ, we see further that for any $k\geq\bar N$,

$$\begin{aligned}
&[\varphi(E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})-\zeta)-\varphi(E(\boldsymbol{x}^{k+1},\boldsymbol{\xi}^{k+1},\boldsymbol{x}^{k})-\zeta)]\cdot\operatorname{dist}(\mathbf{0},\partial E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1}))\\
&\geq\varphi'(E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})-\zeta)\cdot\operatorname{dist}(\mathbf{0},\partial E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1}))\cdot(E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})-E(\boldsymbol{x}^{k+1},\boldsymbol{\xi}^{k+1},\boldsymbol{x}^{k}))\\
&\geq E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})-E(\boldsymbol{x}^{k+1},\boldsymbol{\xi}^{k+1},\boldsymbol{x}^{k}),
\end{aligned}$$

where the last inequality follows from (16) and (9), which states that the sequence $\{E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})\}$ is nonincreasing, thanks to $\sup_k\beta_k<1$. Combining this with (9) and (10), and writing $\Delta_k:=\varphi(E(\boldsymbol{x}^k,\boldsymbol{\xi}^k,\boldsymbol{x}^{k-1})-\zeta)-\varphi(E(\boldsymbol{x}^{k+1},\boldsymbol{\xi}^{k+1},\boldsymbol{x}^{k})-\zeta)$ and $C:=\frac{L}{2}(1-\sup_k\beta_k^2)>0$, we have for any $k\geq\bar N$ that,

$$\|\boldsymbol{x}^k-\boldsymbol{x}^{k-1}\|^2\leq\frac{D}{C}\Delta_k(\|\boldsymbol{x}^k-\boldsymbol{x}^{k-1}\|+\|\boldsymbol{x}^{k-1}-\boldsymbol{x}^{k-2}\|).$$

Therefore, applying the arithmetic mean-geometric mean (AM-GM) inequality, we obtain

$$\begin{aligned}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\| &\le \sqrt{\frac{2D}{C}\Delta_k} \cdot \sqrt{\frac{\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\| + \|\boldsymbol{x}^{k-1} - \boldsymbol{x}^{k-2}\|}{2}} \\ &\le \frac{D}{C}\Delta_k + \frac{1}{4}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\| + \frac{1}{4}\|\boldsymbol{x}^{k-1} - \boldsymbol{x}^{k-2}\|,\end{aligned}$$

which implies that

$$\frac{1}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\| \le \frac{D}{C}\Delta_k + \frac{1}{4}(\|\boldsymbol{x}^{k-1} - \boldsymbol{x}^{k-2}\| - \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|). \tag{17}$$

Summing both sides of (17) from $k = \bar{N}$ to ∞ and noting that $\sum_{k=\bar{N}}^{\infty} \Delta_k \le \varphi(E(\boldsymbol{x}^{\bar{N}}, \boldsymbol{\xi}^{\bar{N}}, \boldsymbol{x}^{\bar{N}-1}) - \zeta)$, we obtain

$$\sum_{k=\bar{N}}^{\infty} \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\| \le \frac{2D}{C}\varphi(E(\boldsymbol{x}^{\bar{N}}, \boldsymbol{\xi}^{\bar{N}}, \boldsymbol{x}^{\bar{N}-1}) - \zeta) + \frac{1}{2}\|\boldsymbol{x}^{\bar{N}-1} - \boldsymbol{x}^{\bar{N}-2}\| < \infty,$$

which implies that the sequence $\{\boldsymbol{x}^k\}$ is convergent. This completes the proof. □

Before closing this section, we would like to point out that, similar to the analysis in [3, Theorem 3.4], one can also establish local convergence rate of the sequence $\{\boldsymbol{x}^k\}$ under the assumptions in Theorem 3.1 and the additional assumption that the function E defined in (7) is a KL function with exponent $\alpha \in [0, 1)$. As an illustration, suppose that the exponent is $\frac{1}{2}$. Then one can show that $\{\boldsymbol{x}^k\}$ converges locally linearly to a stationary point of F in (1). Indeed, according to Proposition 3.1, it holds that $E \equiv \zeta$ for some constant ζ on the compact set Υ. Using this, the KL assumption on E and following the proof of [6, Lemma 6], we conclude that the uniform KL property in (4) holds for E (in place of h) with $\Gamma = \Upsilon$ and $\varphi(s) = cs^{\frac{1}{2}}$ for some $c > 0$. Now, proceed as in the proof of Theorem 3.1 and use $\varphi(s) = cs^{\frac{1}{2}}$ in (16), we have

$$\sqrt{E(\boldsymbol{x}^k, \boldsymbol{\xi}^k, \boldsymbol{x}^{k-1}) - \zeta} \le \frac{c}{2} \cdot \mathrm{dist}\left(0, \partial E(\boldsymbol{x}^k, \boldsymbol{\xi}^k, \boldsymbol{x}^{k-1})\right).$$

Combining this with (10) and (9), we see further that for all sufficiently large k,

$$\begin{aligned}&E(\boldsymbol{x}^{k+1}, \boldsymbol{\xi}^{k+1}, \boldsymbol{x}^k) - \zeta \le E(\boldsymbol{x}^k, \boldsymbol{\xi}^k, \boldsymbol{x}^{k-1}) - \zeta \le \frac{c^2}{4}\mathrm{dist}^2\left(0, \partial E(\boldsymbol{x}^k, \boldsymbol{\xi}^k, \boldsymbol{x}^{k-1})\right) \\ &\le \frac{c^2D^2}{4}(\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\| + \|\boldsymbol{x}^{k-1} - \boldsymbol{x}^{k-2}\|)^2 \le \frac{c^2D^2}{2}(\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2 + \|\boldsymbol{x}^{k-1} - \boldsymbol{x}^{k-2}\|^2) \\ &\le C(E(\boldsymbol{x}^{k-1}, \boldsymbol{\xi}^{k-1}, \boldsymbol{x}^{k-2}) - E(\boldsymbol{x}^{k+1}, \boldsymbol{\xi}^{k+1}, \boldsymbol{x}^k)),\end{aligned}$$

where $C = \frac{c^2D^2}{L(1-\sup_k \beta_k^2)}$. One can then deduce that the sequence $\{E(\boldsymbol{x}^{k+1}, \boldsymbol{\xi}^{k+1}, \boldsymbol{x}^k)\}$ is R-linearly convergent from the above inequality. The R-linear convergence of $\{\boldsymbol{x}^k\}$ now follows from this and (9).

Thus, the KL exponent of E plays an important role in analyzing the local convergence rate of the pDCA$_e$. This is to be contrasted with [40, Theorem 4.3], which analyzed local convergence of the pDCA$_e$ based on the KL exponent of the function $\hat{E}$ in (6) under an additional smoothness assumption on P_2. In the next section, we study a relationship between the KL assumption on E and that on $\hat{E}$; the latter was used in the convergence analysis in [40].

Remark 3.1. *Notice that the classical DCA applied to* (1) *with* $f(\boldsymbol{x}) \equiv 0$ *is a special case of pDCA*$_e$ *(*$\beta_k = 0$*). Thus, one can establish the convergence of the classical DCA without assuming that* P_2 *is smooth using our results.*

4 Connecting various KL assumptions

As discussed at the end of the previous section, the convergence rate of pDCA$_e$ can be analyzed based on the KL exponent of the function E in (7). Specifically, when F in (1) is level-bounded, the sequence $\{\boldsymbol{x}^k\}$ generated by pDCA$_e$ is locally linearly convergent if the exponent is $\frac{1}{2}$. On the other hand, when P_2 satisfies a certain smoothness assumption (see the assumptions on P_2 in Theorem 2.2), a local linear convergence result was established in [40, Theorem 4.3] under a different KL assumption: by assuming that the function $\hat{E}$ in (6) is a KL function with exponent $\frac{1}{2}$. In this section, we study a relationship between these two KL assumptions.

We first prove the following theorem, which studies the KL exponent of a majorant formed from the original function by majorizing the concave part.

Theorem 4.1. *Let $h(\boldsymbol{x}) = Q_1(\boldsymbol{x}) - Q_2(\boldsymbol{A}\boldsymbol{x})$, where Q_1 is proper closed, Q_2 is convex with globally Lipschitz gradient and $\boldsymbol{A}$ is a linear mapping. Suppose that h satisfies the KL property at $\bar{\boldsymbol{x}} \in \operatorname{dom}\partial h$ with exponent $\frac{1}{2}$. Then $H(\boldsymbol{x}, \boldsymbol{y}) = Q_1(\boldsymbol{x}) - \langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}\rangle + Q_2^*(\boldsymbol{y})$ satisfies the KL property at $(\bar{\boldsymbol{x}}, \nabla Q_2(A\bar{\boldsymbol{x}})) \in \operatorname{dom}\partial H$ with exponent $\frac{1}{2}$.*

Proof. Note that it is routine to prove that $(\bar{\boldsymbol{x}}, \nabla Q_2(\boldsymbol{A}\bar{\boldsymbol{x}})) \in \operatorname{dom}\partial H$ and that

$$H(\bar{\boldsymbol{x}}, \nabla Q_2(\boldsymbol{A}\bar{\boldsymbol{x}})) = h(\bar{\boldsymbol{x}}). \tag{18}$$

For any $(\boldsymbol{x}, \boldsymbol{y}) \in \operatorname{dom}\partial H$, let $\tilde{\boldsymbol{u}} \in \partial Q_2^*(\boldsymbol{y})$. Then we have

$$\begin{aligned} H(\boldsymbol{x}, \boldsymbol{y}) &= h(\boldsymbol{x}) + Q_2(\boldsymbol{A}\boldsymbol{x}) + Q_2^*(\boldsymbol{y}) - \langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}\rangle \\ &= h(\boldsymbol{x}) + Q_2(\boldsymbol{A}\boldsymbol{x}) - Q_2(\tilde{\boldsymbol{u}}) + Q_2(\tilde{\boldsymbol{u}}) + Q_2^*(\boldsymbol{y}) - \langle \tilde{\boldsymbol{u}}, \boldsymbol{y}\rangle + \langle \tilde{\boldsymbol{u}} - \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}\rangle \\ &= h(\boldsymbol{x}) + Q_2(\boldsymbol{A}\boldsymbol{x}) - Q_2(\tilde{\boldsymbol{u}}) + \langle \tilde{\boldsymbol{u}} - \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}\rangle \le h(\boldsymbol{x}) + \frac{\ell}{2}\|\tilde{\boldsymbol{u}} - \boldsymbol{A}\boldsymbol{x}\|^2, \end{aligned} \tag{19}$$

where the last equality follows from the fact that $\tilde{\boldsymbol{u}} \in \partial Q_2^*(\boldsymbol{y})$, and the inequality follows from $\tilde{\boldsymbol{u}} \in \partial Q_2^*(\boldsymbol{y})$ (so that $\boldsymbol{y} = \nabla Q_2(\tilde{\boldsymbol{u}})$) and the Lipschitz continuity of ∇Q_2, with ℓ being its Lipschitz continuity modulus. Taking infimum over $\tilde{\boldsymbol{u}} \in \partial Q_2^*(\boldsymbol{y})$, we see from (19) that

$$H(\boldsymbol{x}, \boldsymbol{y}) \le h(\boldsymbol{x}) + \frac{\ell}{2}\operatorname{dist}^2(\boldsymbol{A}\boldsymbol{x}, \partial Q_2^*(\boldsymbol{y})) \tag{20}$$

for any $(\boldsymbol{x}, \boldsymbol{y}) \in \operatorname{dom}\partial H$.

Since h has KL exponent $\frac{1}{2}$ at $\bar{\boldsymbol{x}} \in \operatorname{dom}\partial h = \operatorname{dom}\partial Q_1$, there exist $c \in (0,1)$ and $\epsilon > 0$ so that

$$\operatorname{dist}^2(\boldsymbol{0}, \partial h(\boldsymbol{x})) \ge c(h(\boldsymbol{x}) - h(\bar{\boldsymbol{x}})) \tag{21}$$

whenever $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \le \epsilon$, $\boldsymbol{x} \in \operatorname{dom}\partial Q_1$ and $h(\boldsymbol{x}) < h(\bar{\boldsymbol{x}}) + \epsilon$.[1] Moreover, since Q_2 has globally Lipschitz gradient, we see that ∂Q_2^* is metrically regular at $(\nabla Q_2(\boldsymbol{A}\bar{\boldsymbol{x}}), \boldsymbol{A}\bar{\boldsymbol{x}})$; see [30, Theorem 9.43]. By shrinking ϵ and c if necessary, we conclude that

$$c\|\boldsymbol{A}^\top\boldsymbol{y} - \boldsymbol{A}^\top\nabla Q_2(\boldsymbol{A}\boldsymbol{x})\| \le c\|\boldsymbol{A}^\top\|\|\boldsymbol{y} - \nabla Q_2(\boldsymbol{A}\boldsymbol{x})\| \le \operatorname{dist}(\boldsymbol{A}\boldsymbol{x}, \partial Q_2^*(\boldsymbol{y})) \tag{22}$$

whenever $\max\{\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|, \|\boldsymbol{y} - \nabla Q_2(\boldsymbol{A}\bar{\boldsymbol{x}})\|\} \le \epsilon$.

Now, consider any $(\boldsymbol{x}, \boldsymbol{y}) \in \operatorname{dom}\partial H$ satisfying $\max\{\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|, \|\boldsymbol{y} - \nabla Q_2(\boldsymbol{A}\bar{\boldsymbol{x}})\|\} \le \epsilon$ and

$$H(\bar{\boldsymbol{x}}, \nabla Q_2(\boldsymbol{A}\bar{\boldsymbol{x}})) \le H(\boldsymbol{x}, \boldsymbol{y}) < H(\bar{\boldsymbol{x}}, \nabla Q_2(\boldsymbol{A}\bar{\boldsymbol{x}})) + \epsilon.$$

Then we have for any such $(\boldsymbol{x}, \boldsymbol{y})$ that

$$\boldsymbol{x} \in \operatorname{dom}\partial Q_1, \quad \boldsymbol{y} \in \operatorname{dom}\partial Q_2^* \quad \text{and} \quad h(\bar{\boldsymbol{x}}) + \epsilon = H(\bar{\boldsymbol{x}}, \nabla Q_2(\boldsymbol{A}\bar{\boldsymbol{x}})) + \epsilon > H(\boldsymbol{x}, \boldsymbol{y}) \ge h(\boldsymbol{x}), \tag{23}$$

[1] The requirement $h(\bar{\boldsymbol{x}}) < h(\boldsymbol{x})$ is dropped because (21) holds trivially when $h(\bar{\boldsymbol{x}}) \ge h(\boldsymbol{x})$.

where, in the third relation, the first equality is due to (18) and the last inequality is a consequence of Young's inequality. Furthermore, for any such $(\boldsymbol{x}, \boldsymbol{y})$, we have

$$
\begin{aligned}
&3\,\mathrm{dist}^2(\mathbf{0}, \partial H(\boldsymbol{x}, \boldsymbol{y})) = 3\,\mathrm{dist}^2\left(\mathbf{0}, \begin{bmatrix} -\boldsymbol{A}^\top \boldsymbol{y} + \partial Q_1(\boldsymbol{x}) \\ -\boldsymbol{A}\boldsymbol{x} + \partial Q_2^*(\boldsymbol{y}) \end{bmatrix}\right) = 3\,\mathrm{dist}^2(\boldsymbol{A}^\top \boldsymbol{y}, \partial Q_1(\boldsymbol{x})) + 3\,\mathrm{dist}^2(\boldsymbol{A}\boldsymbol{x}, \partial Q_2^*(\boldsymbol{y}))\\
&\geq 2\,\mathrm{dist}^2(\boldsymbol{A}^\top \boldsymbol{y}, \partial Q_1(\boldsymbol{x})) + 2\,\mathrm{dist}^2(\boldsymbol{A}\boldsymbol{x}, \partial Q_2^*(\boldsymbol{y})) + \mathrm{dist}^2(\boldsymbol{A}\boldsymbol{x}, \partial Q_2^*(\boldsymbol{y}))\\
&\geq c^2[2\,\mathrm{dist}^2(\boldsymbol{A}^\top \boldsymbol{y}, \partial Q_1(\boldsymbol{x})) + 2\|\boldsymbol{A}^\top \boldsymbol{y} - \boldsymbol{A}^\top \nabla Q_2(\boldsymbol{A}\boldsymbol{x})\|^2 + \mathrm{dist}^2(\boldsymbol{A}\boldsymbol{x}, \partial Q_2^*(\boldsymbol{y}))]\\
&\geq c^2\left[\mathrm{dist}^2(\mathbf{0}, \partial h(\boldsymbol{x})) + \mathrm{dist}^2(\boldsymbol{A}\boldsymbol{x}, \partial Q_2^*(\boldsymbol{y}))\right] \geq c'\left[\frac{1}{c}\mathrm{dist}^2(\mathbf{0}, \partial h(\boldsymbol{x})) + \frac{\ell}{2}\mathrm{dist}^2(\boldsymbol{A}\boldsymbol{x}, \partial Q_2^*(\boldsymbol{y}))\right]\\
&\geq c'\left[h(\boldsymbol{x}) - h(\bar{\boldsymbol{x}}) + \frac{\ell}{2}\mathrm{dist}^2(\boldsymbol{A}\boldsymbol{x}, \partial Q_2^*(\boldsymbol{y}))\right] \geq c'\,[H(\boldsymbol{x}, \boldsymbol{y}) - h(\bar{\boldsymbol{x}})] = c'\,[H(\boldsymbol{x}, \boldsymbol{y}) - H(\bar{\boldsymbol{x}}, \nabla Q_2(\boldsymbol{A}\bar{\boldsymbol{x}}))]
\end{aligned}
$$

for $c' := \frac{c^2}{(\frac{1}{c}+\frac{\ell}{2})}$, where the second inequality follows from (22) and the fact that $c < 1$, the third inequality follows from the relation $2(a^2 + b^2) \geq (a + b)^2$ for $a = \mathrm{dist}(\boldsymbol{A}^\top \boldsymbol{y}, \partial Q_1(\boldsymbol{x}))$ and $b = \|\boldsymbol{A}^\top \boldsymbol{y} - \boldsymbol{A}^\top \nabla Q_2(\boldsymbol{A}\boldsymbol{x})\|$, the triangle inequality and the definition of h, the second last inequality follows from (21) and (23), while the last inequality follows from (20). The last equality is due to (18). This completes the proof. □

We are ready to prove the main theorem in this section, which is now an easy corollary of Theorem 4.1. The first conclusion studies a relationship between the KL assumption used in our analysis and the one used in the analysis in [40], while the second conclusion shows that one may deduce the KL exponent of the function E in (7) directly from that of the original objective function F in (1).

Theorem 4.2. *Let F, $\hat{E}$ and E be defined in* (1), (6) *and* (7) *respectively. Suppose in addition that P_2 has globally Lipschitz gradient. Then the following statements hold:*

(i) *If $\hat{E}$ is a KL function with exponent $\frac{1}{2}$, then E is a KL function with exponent $\frac{1}{2}$.*

(ii) *If F is a KL function with exponent $\frac{1}{2}$, then E is a KL function with exponent $\frac{1}{2}$.*

Proof. We first prove (i). Recall from [21, Lemma 2.1] that it suffices to prove that E satisfies the KL property with exponent $\frac{1}{2}$ at all points $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w})$ satisfying $\mathbf{0} \in \partial E(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w})$. To this end, let $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}, \bar{\boldsymbol{w}})$ satisfy $\mathbf{0} \in \partial E(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}, \bar{\boldsymbol{w}})$. Then we obtain from the definition of E that

$$\mathbf{0} \in \nabla f(\bar{\boldsymbol{x}}) + \partial P_1(\bar{\boldsymbol{x}}) - \bar{\boldsymbol{y}} + L(\bar{\boldsymbol{x}} - \bar{\boldsymbol{w}}), \quad \bar{\boldsymbol{x}} \in \partial P_2^*(\bar{\boldsymbol{y}}), \quad \bar{\boldsymbol{x}} = \bar{\boldsymbol{w}}. \tag{24}$$

Plugging the second and the third relations above into the first relation gives

$$\mathbf{0} \in \nabla f(\bar{\boldsymbol{x}}) + \partial P_1(\bar{\boldsymbol{x}}) - \nabla P_2(\bar{\boldsymbol{x}}).$$

This further implies $\mathbf{0} \in \partial \hat{E}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}})$, and hence $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}}) \in \mathrm{dom}\,\partial \hat{E}$. Thus, by assumption, the function $\hat{E}$ satisfies the KL property with exponent $\frac{1}{2}$ at $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}})$. Since

$$\hat{E}(\boldsymbol{x}, \boldsymbol{w}) = f(\boldsymbol{x}) + P_1(\boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{w}\|^2 - P_2\left(\begin{bmatrix} \boldsymbol{I} & \mathbf{0} \end{bmatrix}\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{w} \end{bmatrix}\right),$$

we conclude immediately from Theorem 4.1 that

$$E(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}) = f(\boldsymbol{x}) + P_1(\boldsymbol{x}) - \langle \boldsymbol{x}, \boldsymbol{y} \rangle + P_2^*(\boldsymbol{y}) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{w}\|^2$$

satisfies the KL property with exponent $\frac{1}{2}$ at $(\bar{\boldsymbol{x}}, \nabla P_2(\bar{\boldsymbol{x}}), \bar{\boldsymbol{x}})$, which is just $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}, \bar{\boldsymbol{w}})$ in view of the second and third relations in (24) and the smoothness of P_2. This proves (i).

We now prove (ii). In view of (i) and [21, Lemma 2.1], it suffices to show that $\hat{E}$ satisfies the KL property with exponent $\frac{1}{2}$ at all points $(\boldsymbol{x}, \boldsymbol{y})$ satisfying $\mathbf{0} \in \partial \hat{E}(\boldsymbol{x}, \boldsymbol{y})$. To this end, let $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{w}})$ satisfy $\mathbf{0} \in \partial \hat{E}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{w}})$. Then we see from the definition of $\hat{E}$ that

$$\mathbf{0} \in \nabla f(\bar{\boldsymbol{x}}) + \partial P_1(\bar{\boldsymbol{x}}) - \nabla P_2(\bar{\boldsymbol{x}}) + L(\bar{\boldsymbol{x}} - \bar{\boldsymbol{w}}), \quad \bar{\boldsymbol{x}} = \bar{\boldsymbol{w}}. \tag{25}$$

These relations show that $\mathbf{0} \in \partial F(\bar{\boldsymbol{x}})$, and hence $\bar{\boldsymbol{x}} \in \operatorname{dom} \partial F$. This together with the KL assumption on F and [21, Theorem 3.6] implies that $\hat{E}$ satisfies the KL property at $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}})$, which is just $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{w}})$ in view of the second relation in (25). This completes the proof. □

Before closing this section, we present in the following corollary some specific choices of F in (1) whose corresponding function E defined in (7) is a KL function with exponent $\frac{1}{2}$.

Corollary 4.1. *Let F and E be defined in* (1) *and* (7) *respectively. Suppose that f is quadratic and $P_1 - P_2$ is the* MCP *or* SCAD *function. Then E is a KL function with exponent $\frac{1}{2}$.*

Proof. Notice from [18, Table 1] that for the MCP or SCAD function, P_1 is a positive multiple of the ℓ_1 norm and P_2 is convex with globally Lipschitz gradient. Thus, by Theorem 4.2(ii), it suffices to prove that F is a KL function with exponent $\frac{1}{2}$. Similar to the arguments in [21, Section 5.2], using the special structure of the MCP or SCAD function, one can write

$$F(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i=1}^{n} \min_{1 \le \ell \le m_i} \left\{ f_{i,\ell}(x_i) + \delta_{C_{i,\ell}}(x_i) \right\} = \min_{j \in \mathcal{J}} \left\{ f(\boldsymbol{x}) + \sum_{i=1}^{n} \left[f_{i,j_i}(x_i) + \delta_{C_{i,j_i}}(x_i) \right] \right\},$$

where $C_{i,\ell}$ are closed intervals, $f_{i,\ell}$ are quadratic (or linear) functions, $1 \le \ell \le m_i$, $1 \le i \le n$, and $\mathcal{J} = \{(j_1, \ldots, j_n) \in \mathbb{N}^n : 1 \le j_i \le m_i \ \forall\, i\}$. Notice also that F is a continuous function. Thus, by [21, Corollary 5.2], we see that F is a KL function with exponent $\frac{1}{2}$. This completes the proof. □

5 Application of pDCA$_e$ to simultaneous sparse recovery and outlier detection

Like the problem of sparse learning/recovery discussed in the introduction, the problem of outlier detection is another classical topic in statistics and signal processing. In particular, it has been extensively studied in the area of machine learning. In this context, outliers refer to observations that are somehow statistically different from the majority of the training instances. On the other hand, in signal processing, outlier detection problems naturally arise when the signals transmitted are contaminated by both Gaussian noise and electromyographic noise: the latter exhibits impulsive behavior and results in extreme measurements/outliers; see, for example, [29].

In this section, we will present a nonconvex optimization model incorporating *both* sparse learning/recovery and outlier detection, and discuss how it can be solved by the pDCA$_e$. This is not the first work combining sparse learning/recovery and outlier detection. For instance, there is a huge literature on *robust* compressed sensing, which uses the ℓ_1 regularizer to identify outliers and recover the underlying sparse signal; see [10] and the references therein. As for statistical learning, papers such as [1, 20, 25] already studied such combined models, but their algorithms are simple search algorithms through the space of possible feature subsets and/or the space of possible sample subsets. Recently, the papers [23, 35] studied nonconvex-regularized robust regression models based on M-estimators. They mainly studied theoretical properties (e.g., consistency, breakdown point) of the proposed models and the following disadvantages were left in their algorithms: Smucler and Yohai's algorithm [35] is only for the ℓ_1 regularizer, and Loh's algorithm, which is based on composite gradient descent [23], requires a carefully-chosen initial solution and does not have a global convergence guarantee.

5.1 Simultaneous sparse recovery and outlier detection

In this section, as motivations, we present two concrete scenarios where the problem of simultaneous sparse recovery and outlier detection arises: robust compressed sensing in signal processing and least trimmed squares regression with variable selection in statistics.

5.1.1 Robust compressed sensing

In compressed sensing, an original sparse or approximately sparse signal in high dimension is compressed and then transmitted via some channels. The task is to recover the original high dimensional signal from the relatively lower dimensional possibly noisy received signal. This problem is NP hard in general; see [26].

When there is no noise in the transmission, the recovery problem can be shown to be equivalent to an ℓ_1 minimization problem under some additional assumptions; see, for example, [7, 14]. Recently, various nonconvex models have also been proposed for recovering the underlying sparse/approximately sparse signal; see, for example, [9, 12, 13]. These models empirically require fewer measurements than their convex counterparts for recovering signals of the same sparsity level.

While the noiseless scenario leads to a theory of exact recovery, in practice, the received signals are noisy. This latter scenario has also been extensively studied in the literature, with Gaussian measurement noise being the typical noise model; see, for example, [8, 33]. However, in certain compressed sensing system, the signals can be corrupted by *both* Gaussian noise and electromyographic noise: the latter exhibits impulsive behavior and results in extreme measurements/outliers [29]. In the literature, the following model was proposed for handling noise and outliers simultaneously, which makes use of ℓ_1 regularizer for both sparse recovery and outlier detection; see the recent exposition [10] and references therein:

$$\begin{array}{cl} \min\limits_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{z}\in\mathbb{R}^m} & \tau\|\boldsymbol{x}\|_1 + \|\boldsymbol{z}\|_1 \\ \text{s.t.} & \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} - \boldsymbol{z}\| \le \epsilon; \end{array}$$

here, $\boldsymbol{A}$ is the sensing matrix, $\boldsymbol{b}$ is the possibly noisy received signal, and $\tau > 0$ and $\epsilon > 0$ are parameters controlling the sparsity in $\boldsymbol{x}$ and the allowable noise level, respectively. In this section, we describe an alternative model that can incorporate some prior knowledge of the number of outliers. In our model, instead of relying on the ℓ_1 norm for detecting outliers, we employ the ℓ_0 norm directly, assuming a rough (upper) estimation r of the number of outliers. We also allow the use of possibly nonconvex regularizers P for inducing sparsity in the signal when there is no prior knowledge of the sparsity level of $\boldsymbol{x}$: these kinds of regularizers have been widely used in the literature and have been shown empirically to work better than convex regularizers; see, for example, [9, 12, 13, 16, 33]. Specifically, our model takes the following form:

$$\begin{array}{cl} \min\limits_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{z}\in\mathbb{R}^m} & \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} - \boldsymbol{z}\|^2 + P(\boldsymbol{x}) \\ \text{s.t.} & \|\boldsymbol{z}\|_0 \le r. \end{array} \tag{26}$$

Notice that at optimality, at most r number of z_i's will be nonzero and equal to the corresponding $[\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}]_i$, zeroing out the corresponding terms in the least squares. Thus, once the nonzero entries of $\boldsymbol{z}$ at optimality are identified, the problem reduces to a standard compressed sensing problem with at least $m - r$ measurements: for this class of problem, (approximate) recovery of the original sparse signal is possible if there are sufficient measurements, assuming the sensing matrix is generated according to certain distributions [7, 8]. This means that one only needs a reasonable *upper bound* on the number of outliers so that $m - r$ is not too small in order to recover the original approximately sparse signal, assuming a random sensing matrix and that the outliers are successfully detected. This is in contrast to some ℓ_0 based approaches such as the iterative hard thresholding (IHT) for compressed sensing [5], where the *exact knowledge* of the sparsity level is needed for recovering the signal.

5.1.2 Least trimmed squares regression with variable selection

In statistics, suppose that we have data samples $\{\boldsymbol{a}_i, b_i\}_{i=1}^m$, where $\boldsymbol{a}_i \in \mathbb{R}^n, b_i \in \mathbb{R}$, and suppose that some data samples are from anomalous observations; such observations are called outliers. These outliers may be caused by mechanical faults, human errors, instrument errors, changes in system behaviour, etc. We need to identify and remove the outliers to improve the prediction performance of the regression model. In this section, we consider the problem of simultaneously identifying the outliers in the set of samples $\{\boldsymbol{a}_i, b_i\}_{i=1}^m$ and recovering a vector $\boldsymbol{x}^* \in \mathbb{R}^n$ using $\{\boldsymbol{a}_i, b_i\}_{i=1}^m$ but the outliers.

Statisticians and data analysts have been searching for regressors which are not affected by outliers, i.e., the regressors that are robust with respect to outliers. Least trimmed squares (LTS) regression [31, 32] is popular as a robust regression model and can be formulated as a nonlinear mixed zero-one integer optimization problem (see (2.1.1) in [17]):

$$\begin{array}{cl} \displaystyle\min_{\substack{\boldsymbol{s}\in\{0,1\}^m,\\ \boldsymbol{x}\in\mathbb{R}^n}} & \displaystyle\sum_{i=1}^m s_i(\boldsymbol{a}_i^\top \boldsymbol{x} - b_i)^2 \\ \text{s.t.} & \displaystyle\sum_{i=1}^m s_i \geq m - r. \end{array}$$

Note that the problem can be equivalently transformed into

$$\begin{array}{cl} \displaystyle\min_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{z}\in\mathbb{R}^m} & \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} - \boldsymbol{z}\|^2 \\ \text{s.t.} & \|\boldsymbol{z}\|_0 \leq r, \end{array} \tag{27}$$

where $\boldsymbol{A} = \begin{bmatrix}\boldsymbol{a}_1 & \cdots & \boldsymbol{a}_m\end{bmatrix}^\top$ and $\boldsymbol{b} = \begin{bmatrix}b_1 & \cdots & b_m\end{bmatrix}^\top$. The above ℓ_0-norm constrained problem was considered in [36]. More recently, an outlier detection problem using a nonconvex regularizer such as soft and hard thresholdings, SCAD [15], etc., was proposed in [34]. In particular, they did not impose the constraint $\|\boldsymbol{z}\|_0 \leq r$ directly.

Most robust regression models make use of the squared loss function, i.e., $\Psi(\boldsymbol{u}) := \sum_{i=1}^m \psi_i(u_i)$ with $\psi_i(s) = \frac{1}{2}(s - b_i)^2$. Alternatively, one can develop robust regression models based on the following loss functions, which are also commonly used in other branches of statistics:

- quadratic ϵ-insensitive loss: $\psi_i(s) = \frac{1}{2}(|s - b_i| - \epsilon)_+^2$, where $(a)_+ := \max\{0, a\}$, for a given hyperparameter $\epsilon > 0$;
- quantile squared loss: $\psi_i(s) = \frac{\tau}{2}(s - b_i)_+^2 + \frac{(1-\tau)}{2}(-s + b_i)_+^2$ for a given hyperparameter $0 < \tau < 1$;

and $s = \boldsymbol{a}_i^\top \boldsymbol{x}$ is assumed for regression problems. These loss functions can be robustified by incorporating a variable z_i into them in the form of $\psi_i(\boldsymbol{a}_i^\top \boldsymbol{x} - z_i)$ as in the squared-loss model (27). We can also add a regularization term, e.g., a nonconvex regularizer $P(\boldsymbol{x})$ for inducing sparsity, to the robust regression problem (27). This can improve the predictive error of the model by reducing the variability in the estimates of regression coefficients by shrinking the estimates towards zero.

The resulting model takes the following from:

$$\begin{array}{ll}\min\limits_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{z}\in\mathbb{R}^m} & \sum\limits_{i=1}^m \psi_i(\boldsymbol{a}_i^\top\boldsymbol{x}-z_i)+P(\boldsymbol{x})\\ \text{s.t.} & \|\boldsymbol{z}\|_0\le r.\end{array} \tag{28}$$

In the case when ψ_i is the squared loss function, the above model can be naturally referred to as the least trimmed squares regression with variable selection.

5.2 A general model and algorithm

In this section, we present a general model that covers the simultaneous sparse recovery and outlier detection models discussed in Section 5.1 for a large class of nonconvex regularizers, and discuss how the model can be solved by pDCA$_e$.

Specifically, we consider the following model:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{z}\in\mathbb{R}^m} \ \Phi(\boldsymbol{x},\boldsymbol{z}):=\Psi(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{z})+\delta_\Omega(\boldsymbol{z})+\mathcal{J}_1(\boldsymbol{x})-\mathcal{J}_2(\boldsymbol{x}), \tag{29}$$

where $\Psi(\boldsymbol{s}):=\sum_{i=1}^m\psi_i(s_i)$ with $\psi_i:\mathbb{R}\to[0,\infty)$ being convex with Lipschitz continuous gradient whose Lipschitz continuity modulus is L_i, $\boldsymbol{A}\in\mathbb{R}^{m\times n}$, $\Omega=\{\boldsymbol{z}\in\mathbb{R}^m:\|\boldsymbol{z}\|_0\le r\}$ for some positive integer r, $\mathcal{J}_1$ is a proper closed convex function and $\mathcal{J}_2$ is a *continuous* convex function. In addition, we assume that $\operatorname{Arg\,min}\psi_i\neq\emptyset$ for each i and that $\mathcal{J}_1-\mathcal{J}_2$ is level-bounded. One can show that the squared loss function, the quadratic ϵ-insensitive loss and the quantile squared loss function mentioned in Section 5.1.2 satisfy the assumptions on ψ_i. Thus, when the regularizer P in (26) or (28) is level-bounded and can be written as the difference of a proper closed convex function and a continuous convex function, then the corresponding problem is a special case of (29).

In order to apply the pDCA$_e$, we need to derive an explicit DC decomposition of the objective in (29) into the form of (1). To this end, we first note that $\nabla\Psi$ is Lipschitz continuous with a Lipschitz continuity modulus of $L_\Psi:=\max_{1\le i\le m}L_i$. Then we know that $g(\boldsymbol{s}):=\frac{L_\Psi}{2}\|\boldsymbol{s}\|^2-\Psi(\boldsymbol{s})$ is convex and continuously differentiable. Hence,

$$\begin{aligned}&\inf_{\boldsymbol{z}\in\mathbb{R}^m}\Phi(\boldsymbol{x},\boldsymbol{z})=\inf_{\boldsymbol{z}\in\mathbb{R}^m}\Psi(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{z})+\delta_\Omega(\boldsymbol{z})+\mathcal{J}_1(\boldsymbol{x})-\mathcal{J}_2(\boldsymbol{x})\\&=\inf_{\boldsymbol{z}\in\Omega}\left[\frac{L_\Psi}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{z}\|^2-g(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{z})\right]+\mathcal{J}_1(\boldsymbol{x})-\mathcal{J}_2(\boldsymbol{x})\\&=\frac{L_\Psi}{2}\|\boldsymbol{A}\boldsymbol{x}\|^2+\mathcal{J}_1(\boldsymbol{x})-\underbrace{\sup_{\boldsymbol{z}\in\Omega}\left\{L_\Psi\langle\boldsymbol{z},\boldsymbol{A}\boldsymbol{x}\rangle-\frac{L_\Psi}{2}\|\boldsymbol{z}\|^2+g(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{z})\right\}}_{Q(\boldsymbol{x})}-\mathcal{J}_2(\boldsymbol{x}).\end{aligned} \tag{30}$$

Now, notice that for each $\boldsymbol{z}\in\Omega$, the function $\boldsymbol{x}\mapsto L_\Psi\langle\boldsymbol{z},\boldsymbol{A}\boldsymbol{x}\rangle-\frac{L_\Psi}{2}\|\boldsymbol{z}\|^2+g(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{z})$ is convex and Q is the pointwise supremum of these functions. Therefore, Q is a convex function. In addition, one can see from (30) that

$$Q(\boldsymbol{x})=\frac{L_\Psi}{2}\|\boldsymbol{A}\boldsymbol{x}\|^2-\inf_{\boldsymbol{z}\in\Omega}\Psi(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{z}). \tag{31}$$

In particular, Q is a convex function that is finite everywhere, and is hence continuous. Using these observations, we can now rewrite (29) as

$$\min_{\boldsymbol{x}\in\mathbb{R}^n}\frac{L_\Psi}{2}\|\boldsymbol{A}\boldsymbol{x}\|^2+P_1(\boldsymbol{x})-P_2(\boldsymbol{x}), \tag{32}$$

where $P_1(\boldsymbol{x}) := \mathcal{J}_1(\boldsymbol{x})$, and $P_2(\boldsymbol{x}) := Q(\boldsymbol{x}) + \mathcal{J}_2(\boldsymbol{x})$ is a continuous convex function. This problem is in the form of (1), and its objective is level-bounded in view of (30), the level-boundedness of $\mathcal{J}_1 - \mathcal{J}_2$ and the nonnegativity of Ψ. Hence, the pDCA$_e$ is applicable for solving it.

In each iteration of the pDCA$_e$, one has to compute the proximal mapping of $P_1 = \mathcal{J}_1$ and a subgradient of $P_2 = Q + \mathcal{J}_2$. Since Q is continuous, it is well known that $\partial P_2(\boldsymbol{x}) = \partial Q(\boldsymbol{x}) + \partial \mathcal{J}_2(\boldsymbol{x})$ for all $\boldsymbol{x}$. The ease of computation of the proximal mapping of $\mathcal{J}_1$ and a subgradient of $\mathcal{J}_2$ depends on the choice of regularizer, while a subgradient of Q at $\boldsymbol{x}$ is readily computable using the observation that for any $\bar{\boldsymbol{z}} \in \operatorname{Arg\,min}_{\boldsymbol{z}\in\Omega} \Psi(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{z})$, we have

$$L_\Psi \boldsymbol{A}^\top \bar{\boldsymbol{z}} + \boldsymbol{A}^\top \nabla g(\boldsymbol{A}\boldsymbol{x} - \bar{\boldsymbol{z}}) = L_\Psi \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}^\top \nabla \Psi(\boldsymbol{A}\boldsymbol{x} - \bar{\boldsymbol{z}}) \in \partial Q(\boldsymbol{x}); \tag{33}$$

this inclusion follows immediately from the definition of Q in (30) and the definition of convex subdifferential.

We are now ready to present the pDCA$_e$ for solving (32) (and hence (29)) as Algorithm 2 below. Notice that this algorithm is just Algorithm 1 applied to (32) with a subgradient of Q computed

---

Algorithm 2 pDCA$_e$ for (29):

Input: $\boldsymbol{x}^0 \in \operatorname{dom} \mathcal{J}_1$, $\{\beta_k\} \subseteq [0,1)$ with $\sup_k \beta_k < 1$ and $L \geq L_\Psi \lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})$. Set $\boldsymbol{x}^{-1} = \boldsymbol{x}^0$.

for $k = 0, 1, 2, \ldots$

take any $\boldsymbol{\eta}^{k+1} \in \partial \mathcal{J}_2(\boldsymbol{x}^k)$, $\boldsymbol{z}^{k+1} \in \operatorname*{Arg\,min}_{\boldsymbol{z}\in\Omega} \Psi(\boldsymbol{A}\boldsymbol{x}^k - \boldsymbol{z})$ and set

$$\begin{cases} \boldsymbol{u}^k = \boldsymbol{x}^k + \beta_k(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}), \\ \boldsymbol{v}^k = \boldsymbol{A}^\top \nabla \Psi(\boldsymbol{A}\boldsymbol{x}^k - \boldsymbol{z}^{k+1}) + L_\Psi \boldsymbol{A}^\top \boldsymbol{A}(\boldsymbol{u}^k - \boldsymbol{x}^k) - \boldsymbol{\eta}^{k+1}, \\ \boldsymbol{x}^{k+1} = \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ \langle \boldsymbol{v}^k, \boldsymbol{x} \rangle + \dfrac{L}{2}\|\boldsymbol{x} - \boldsymbol{u}^k\|^2 + \mathcal{J}_1(\boldsymbol{x}) \right\}. \end{cases}$$

end for

---

as in (33) in each step. The following lemma gives a closed-form solution for the $\boldsymbol{z}$-update in Algorithm 2, and is an immediate corollary of [24, Proposition 3.1].

Lemma 5.1. *Fix any $\tilde{\boldsymbol{z}} \in \operatorname{Arg\,min}_{\boldsymbol{z}\in\mathbb{R}^m} \Psi(\boldsymbol{z}) = \operatorname{Arg\,min}_{\boldsymbol{z}\in\mathbb{R}^m} \sum_{i=1}^m \psi_i(z_i)$ and let $\tilde{\boldsymbol{z}}^k = \boldsymbol{A}\boldsymbol{x}^k - \tilde{\boldsymbol{z}}$. Let $I^* \subseteq \{1, \ldots, m\}$ be an index set corresponding to any r largest values of $\{\psi_i([\boldsymbol{A}\boldsymbol{x}^k]_i) - \psi_i(\tilde{z}_i)\}_{i=1}^m$ and set*

$$z_i^{k+1} = \begin{cases} \tilde{z}_i^k & \text{if } i \in I^*, \\ 0 & \text{otherwise.} \end{cases}$$

Then we have

$$\boldsymbol{z}^{k+1} \in \operatorname*{Arg\,min}_{\boldsymbol{z}\in\Omega} \Psi(\boldsymbol{A}\boldsymbol{x}^k - \boldsymbol{z}).$$

In the last theorem of this section, we show that for problem (32) (equivalently, (29)) with many commonly used loss functions ψ_i and regularizers $\mathcal{J}_1 - \mathcal{J}_2$, the corresponding potential function E is a KL function with exponent $\frac{1}{2}$. This together with the discussion at the end of Section 3 reveals that the pDCA$_e$ is locally linearly convergent when applied to these models. In the proof below, for notational simplicity, for a positive integer m, we let S_m denote the set of all possible permutations of $\{1, \ldots, m\}$.

Theorem 5.1. *Let Q be given in (31), with ψ_i taking one of the following forms:*

(i) *squared loss:* $\psi_i(s) = \frac{1}{2}(s - b_i)^2$, $b_i \in \mathbb{R}$;

(ii) *squared hinge loss:* $\psi_i(s) = \frac{1}{2}(1 - b_i s)_+^2$, $b_i \in \{1, -1\}$;

(iii) *quadratic ϵ-insensitive loss:* $\psi_i(s) = \frac{1}{2}(|s - b_i| - \epsilon)_+^2$, $\epsilon > 0$ *and* $b_i \in \mathbb{R}$;

(iv) *quantile squared loss:* $\psi_i(s) = \frac{\tau}{2}(s - b_i)_+^2 + \frac{(1-\tau)}{2}(-s + b_i)_+^2$, $0 < \tau < 1$ *and* $b_i \in \mathbb{R}$.

Then Q is a convex piecewise linear-quadratic function. Suppose in addition that $\mathcal{J}_1$ and $\mathcal{J}_2$ are convex piecewise linear-quadratic functions. Then the function E in (7) *corresponding to* (32) *is a KL function with exponent $\frac{1}{2}$.*

Proof. We first prove that Q is convex piecewise linear-quadratic when ψ_i is chosen as one of the four loss functions. We start with (i). Clearly, $L_\Psi = 1$ and we have from (31) that

$$Q(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}\|^2 - \inf_{\boldsymbol{z}\in\Omega} \Psi(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{z}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}\|^2 - \inf_{\boldsymbol{z}\in\Omega} \left\{ \sum_{i=1}^{m} \frac{1}{2}([\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}]_i - z_i)^2 \right\}. \tag{34}$$

Let $\mathcal{I}$ be an index set corresponding to any r largest entries of $\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}$ in magnitude. We then see from Lemma 5.1 that if

$$z_i^* = \begin{cases} [\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}]_i & \text{if } i \in \mathcal{I}, \\ 0 & \text{otherwise,} \end{cases}$$

then $\boldsymbol{z}^*$ attains the infimum in (34). Thus, we have

$$Q(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}\|^2 - \sum_{i=r+1}^{m} \frac{1}{2}\left([\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}]_{[i]}\right)^2,$$

where $[\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}]_{[i]}$ denotes the ith largest entry of $\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}$ in magnitude. Notice that for each fixed permutation $\sigma \in S_m$, the set

$$\Omega_\sigma = \left\{\boldsymbol{x} : \left|[\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}]_{\sigma(1)}\right| \geq \left|[\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}]_{\sigma(2)}\right| \geq \ldots \geq \left|[\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}]_{\sigma(m)}\right|\right\}$$

is a union of finitely many polyhedra and the restriction of Q onto Ω_σ is a quadratic function. Moreover, $\bigcup_{\sigma\in S_m} \Omega_\sigma = \mathbb{R}^n$. Thus, Q is a piecewise linear-quadratic function when ψ_i takes the form in (i).

Then we consider case (ii). Again, $L_\Psi = 1$, and we have from (31), $b_i \in \{1, -1\}$ and Lemma 5.1 that

$$\begin{aligned} Q(\boldsymbol{x}) &= \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}\|^2 - \inf_{\boldsymbol{z}\in\Omega} \Psi(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{z}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}\|^2 - \inf_{\boldsymbol{z}\in\Omega} \left\{ \sum_{i=1}^{m} \frac{1}{2}(1 - b_i[\boldsymbol{A}\boldsymbol{x} - \boldsymbol{z}]_i)_+^2 \right\} \\ &= \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}\|^2 - \inf_{\boldsymbol{z}\in\Omega} \left\{ \sum_{i=1}^{m} \frac{1}{2}(1 - b_i[\boldsymbol{A}\boldsymbol{x}]_i - z_i)_+^2 \right\} = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}\|^2 - \sum_{i=r+1}^{m} \frac{1}{2}\left([\boldsymbol{e} - \boldsymbol{b} \circ \boldsymbol{A}\boldsymbol{x}]_{(i)}\right)_+^2 \end{aligned}$$

where $[\boldsymbol{e} - \boldsymbol{b} \circ \boldsymbol{A}\boldsymbol{x}]_{(i)}$ the ith largest entry of $\boldsymbol{e} - \boldsymbol{b} \circ \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{e} \in \mathbb{R}^m$ is the vector of all ones. Note that for each fixed permutation $\sigma \in S_m$, the set

$$\Omega_\sigma = \left\{\boldsymbol{x} : [\boldsymbol{e} - \boldsymbol{b} \circ \boldsymbol{A}\boldsymbol{x}]_{\sigma(1)} \geq [\boldsymbol{e} - \boldsymbol{b} \circ \boldsymbol{A}\boldsymbol{x}]_{\sigma(2)} \geq \ldots \geq [\boldsymbol{e} - \boldsymbol{b} \circ \boldsymbol{A}\boldsymbol{x}]_{\sigma(m)}\right\}$$

is a polyhedron and the restriction of Q onto Ω_σ is a piecewise linear-quadratic function. Moreover, $\bigcup_{\sigma\in S_m} \Omega_\sigma = \mathbb{R}^n$. Thus, Q is a piecewise linear-quadratic function when ψ_i takes the form in (ii).

Next, we turn to case (iii). Again, $L_\Psi = 1$, and we have from (31) and Lemma 5.1 that

$$Q(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}\|^2 - \inf_{\boldsymbol{z}\in\Omega}\left\{\frac{1}{2}\sum_{i=1}^m(|[\boldsymbol{A}\boldsymbol{x}-\boldsymbol{z}]_i - b_i| - \epsilon)_+^2\right\} = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}\|^2 - \sum_{i=r+1}^m \frac{1}{2}\left(\left|[\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}]_{[i]}\right| - \epsilon\right)_+^2,$$

where $[\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}]_{[i]}$ denotes the ith largest entry of $\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}$ in magnitude. Using a similar argument as above, one can see that Q is a piecewise linear-quadratic function.

Finally, we consider case (iv). Define $w_i = \sqrt{\frac{\tau}{2}}([\boldsymbol{A}\boldsymbol{x}]_i - b_i)_+ + \sqrt{\frac{1-\tau}{2}}(-[\boldsymbol{A}\boldsymbol{x}]_i + b_i)_+$, $i = 1, \ldots, m$. Then

$$w_i^2 = \frac{\tau}{2}([\boldsymbol{A}\boldsymbol{x}]_i - b_i)_+^2 + \frac{1-\tau}{2}(-[\boldsymbol{A}\boldsymbol{x}]_i + b_i)_+^2,$$

and for all i, j, it holds that $w_i \geq w_j$ if and only if $w_i^2 \geq w_j^2$. Since we can take $L_\Psi = 1$, we have from these and (31) that

$$\begin{aligned} Q(\boldsymbol{x}) &= \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}\|^2 - \inf_{\boldsymbol{z}\in\Omega}\left\{\sum_{i=1}^m \frac{\tau}{2}([\boldsymbol{A}\boldsymbol{x}]_i - z_i - b_i)_+^2 + \frac{1-\tau}{2}(-[\boldsymbol{A}\boldsymbol{x}]_i + z_i + b_i)_+^2\right\} \\ &= \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}\|^2 - \sum_{i=r+1}^m (w_{\{i\}})^2, \end{aligned}$$

where $w_{\{i\}}$ denotes the ith largest element of $\{w_i\}_{i=1,\ldots,m}$. For each fixed permutation $\sigma \in S_m$, we define a set

$$\Omega_\sigma := \left\{\boldsymbol{x} : w_{\sigma(1)} \geq w_{\sigma(2)} \geq \ldots \geq w_{\sigma(m)}\right\}.$$

Notice that the restriction of Q onto Ω_σ is a piecewise linear-quadratic function. Moreover, Ω_σ can be written as a union of finitely many polyhedra and $\bigcup_{\sigma\in S_m}\Omega_\sigma = \mathbb{R}^n$. Thus, Q is a piecewise linear-quadratic function.

Finally, we show that E is a KL function with exponent $\frac{1}{2}$ under the additional assumption that $\mathcal{J}_1$ and $\mathcal{J}_2$ are convex piecewise linear-quadratic functions. Notice from (32) and (7) that

$$E(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}\|^2 + \mathcal{J}_1(\boldsymbol{x}) - \langle\boldsymbol{x},\boldsymbol{y}\rangle + (Q+\mathcal{J}_2)^*(\boldsymbol{y}) + \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{w}\|^2. \tag{35}$$

Since Q and $\mathcal{J}_2$ are convex piecewise linear-quadratic functions, we know from [30, Exercise 10.22] and [30, Theorem 11.14] that $(Q+\mathcal{J}_2)^*$ is also a piecewise linear-quadratic function. Hence, $\mathcal{J}_1(\boldsymbol{x}) + (Q+\mathcal{J}_2)^*(\boldsymbol{y})$ is also a piecewise linear-quadratic functions and can be written as

$$\mathcal{J}_1(\boldsymbol{x}) + (Q+\mathcal{J}_2)^*(\boldsymbol{y}) = \min_{1\leq i\leq M}\left\{g_i(\boldsymbol{x},\boldsymbol{y}) + \delta_{C_i}(\boldsymbol{x},\boldsymbol{y})\right\},$$

where $M > 0$ is an integer, g_i are quadratic functions and C_i are polyhedra. Then we have

$$E(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w}) = \min_{1\leq i\leq M}\left\{\frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}\|^2 - \langle\boldsymbol{x},\boldsymbol{y}\rangle + \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{w}\|^2 + g_i(\boldsymbol{x},\boldsymbol{y}) + \delta_{C_i}(\boldsymbol{x},\boldsymbol{y})\right\}.$$

Moreover, this function is continuous in its domain because, according to (35), it is the sum of piecewise linear-quadratic functions which are continuous in their domains [30, Proposition 10.21]. Thus, by [21, Corollary 5.2], this function is a KL function with exponent $\frac{1}{2}$. This completes the proof. □

6 Numerical simulations

In this section, we perform numerical experiments to explore the performance of pDCA$_e$ in some specific simultaneous sparse recovery and outlier detection problems. All experiments are performed in Matlab R2015b on a 64-bit PC with an Intel(R) Core(TM) i7-4790 CPU (3.60GHz) and 32GB of RAM.

We consider the following special case of (29) with the least trimmed squares loss function and the Truncated ℓ_1 regularizer:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{z}\in\mathbb{R}^m} \ \Phi_{\rm trc}(\boldsymbol{x},\boldsymbol{z}) := \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{z}-\boldsymbol{b}\|^2 + \delta_\Omega(\boldsymbol{z}) + \lambda\|\boldsymbol{x}\|_1 - \lambda\mu\sum_{i=1}^{p}|x_{[i]}|, \tag{36}$$

where $\boldsymbol{A}\in\mathbb{R}^{m\times n}$, $\boldsymbol{b}\in\mathbb{R}^m$, $\Omega=\{\boldsymbol{z}\in\mathbb{R}^m : \|\boldsymbol{z}\|_0\le r\}$, $\mu\in(0,1)$, $\lambda>0$ is the regularization parameter, $p<n$ is a positive integer and $x_{[i]}$ denotes the ith largest entry of $\boldsymbol{x}$ in magnitude. One can see that $\Phi_{\rm trc}$ takes the form of (29), where $\psi_i(s)=\frac{1}{2}(s-b_i)^2$, $\mathcal{J}_1(\boldsymbol{x})=\lambda\|\boldsymbol{x}\|_1$ and $\mathcal{J}_2(\boldsymbol{x})=\lambda\mu\sum_{i=1}^p|x_{[i]}|$ with $\mathcal{J}_1-\mathcal{J}_2$ being level-bounded. In this case, $L_\Psi=1$, and we can rewrite (36) in the form (1) (see also (32)):

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \ F_{\rm trc}(\boldsymbol{x}) := \underbrace{\frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}\|^2}_{f(\boldsymbol{x})} + \underbrace{\lambda\|\boldsymbol{x}\|_1}_{P_1(\boldsymbol{x})} - \underbrace{\left(\lambda\mu\sum_{i=1}^{p}|x_{[i]}| + Q(\boldsymbol{x})\right)}_{P_2(\boldsymbol{x})}, \tag{37}$$

where Q is defined in (31).

Notice that $\mathcal{J}_1$ and $\mathcal{J}_2$ are piecewise linear-quadratic functions. By Theorem 5.1, we see that for the function $F_{\rm trc}$ given in (37), the corresponding function E in (7) is a KL function with exponent $\frac{1}{2}$. Thus, we conclude from Theorem 3.1 and the discussion at the end of Section 3 that the sequence $\{\boldsymbol{x}^k\}$ generated by pDCA$_e$ for solving (37) converges locally linearly to a stationary point of $F_{\rm trc}$ in (37).

In our experiments below, we compare pDCA$_e$ with NPG$_{\rm major}$ [22] for solving (37). We discuss the implementation details below. For the ease of exposition, we introduce an auxiliary function:

$$\Xi(\boldsymbol{x},\boldsymbol{y}) = f(\boldsymbol{x}) + P_1(\boldsymbol{x}) - \langle\boldsymbol{x},\boldsymbol{y}\rangle + P_2^*(\boldsymbol{y}),$$

where f, P_1 and P_2 are defined in (37). Notice that $\boldsymbol{0}\in\partial\Xi(\bar{\boldsymbol{x}},\bar{\boldsymbol{y}})$ for some $\bar{\boldsymbol{y}}$ if and only if $\bar{\boldsymbol{x}}$ is a stationary point of $F_{\rm trc}$ in (37), i.e., $\boldsymbol{0}\in\nabla f(\bar{\boldsymbol{x}})+\partial P_1(\bar{\boldsymbol{x}})-\partial P_2(\bar{\boldsymbol{x}})$. We will employ Ξ in the design of termination criterion for the algorithms.

pDCA$_e$. We apply Algorithm 1 to (37), with a subgradient of Q computed as in (33) in each step.[2] In our experiments below, as in [40, Section 5], we set $\boldsymbol{x}^0=\boldsymbol{0}$, $L=\lambda_{\max}(\boldsymbol{A}^\top\boldsymbol{A})$ and start with $\theta_{-1}=\theta_0=1$, recursively define for $k\ge 0$ that

$$\beta_k=\theta_k(\theta_{k-1}^{-1}-1) \ \text{ with } \ \theta_{k+1}=\frac{2}{1+\sqrt{1+4/\theta_k^2}}.$$

We then reset $\theta_{-1}=\theta_0=1$ every 200 iterations. To derive a reasonable termination criterion, we first note from the first-order optimality condition of the $\boldsymbol{x}$-update in (5) that

$$-\nabla f(\boldsymbol{u}^{k-1}) + \boldsymbol{\xi}^k - L(\boldsymbol{x}^k-\boldsymbol{u}^{k-1}) \in \partial P_1(\boldsymbol{x}^k).$$

This together with $\boldsymbol{x}^{k-1}\in\partial P_2^*(\boldsymbol{\xi}^k)$ and

$$\partial\Xi(\boldsymbol{x}^k,\boldsymbol{\xi}^k) = \begin{bmatrix}\nabla f(\boldsymbol{x}^k)+\partial P_1(\boldsymbol{x}^k)-\boldsymbol{\xi}^k\\ -\boldsymbol{x}^k+\partial P_2^*(\boldsymbol{\xi}^k)\end{bmatrix}$$

implies that

$$\begin{bmatrix}\nabla f(\boldsymbol{x}^k)-\nabla f(\boldsymbol{u}^{k-1})-L(\boldsymbol{x}^k-\boldsymbol{u}^{k-1})\\ -\boldsymbol{x}^k+\boldsymbol{x}^{k-1}\end{bmatrix}\in\partial\Xi(\boldsymbol{x}^k,\boldsymbol{\xi}^k).$$

[2] As mentioned before, with this choice of subgradient in Algorithm 1, the algorithm is equivalent to Algorithm 2.

Thus, we terminate the algorithm when

$$\sqrt{\left(\sqrt{L}\|\boldsymbol{A}(\boldsymbol{x}^k-\boldsymbol{u}^{k-1})\|+L\|\boldsymbol{x}^k-\boldsymbol{u}^{k-1}\|\right)^2+\|\boldsymbol{x}^k-\boldsymbol{x}^{k-1}\|^2}<10^{-4}\max\{1,\|\boldsymbol{x}^k\|\},$$

so that we have $\operatorname{dist}(0,\partial\Xi(\boldsymbol{x}^k,\boldsymbol{\xi}^k))<10^{-4}\max\{1,\|\boldsymbol{x}^k\|\}$.

NPG$_{\rm major}$. We solve (37) by the NPG$_{\rm major}$ algorithm described in [22, Apendix A, Algorithm 2], which is basically the proximal DCA incorporated with a nonmonotone linesearch scheme. Following the notation there, we apply the method with $h(\boldsymbol{x})=f(\boldsymbol{x})$, $P(\boldsymbol{x})=P_1(\boldsymbol{x})$ and $g(\boldsymbol{x})=P_2(\boldsymbol{x})$, and set $\boldsymbol{x}^0=\boldsymbol{0}$, $\tau=2$, $c=10^{-4}$, $M=4$, $L_0^0=1$, $L_{\min}=10^{-8}$, $L_{\max}=10^8$ and

$$L_k^0=\begin{cases}\min\left\{\max\left\{\frac{{\boldsymbol{s}^k}^\top\boldsymbol{y}^k}{\|\boldsymbol{s}^k\|^2},L_{\min}\right\},L_{\max}\right\} & \text{if } {\boldsymbol{s}^k}^\top\boldsymbol{y}^k\geq 10^{-12},\\ \min\left\{\max\left\{\frac{\bar{L}_{k-1}}{2},L_{\min}\right\},L_{\max}\right\} & \text{otherwise,}\end{cases}$$

for $k\geq 1$; here, $\bar{L}_{k-1}$ is determined in [22, Apendix A, Algorithm 2, Step 2], $\boldsymbol{s}^k=\boldsymbol{x}^k-\boldsymbol{x}^{k-1}$ and $\boldsymbol{y}^k=\boldsymbol{A}^\top\left(\boldsymbol{A}\boldsymbol{x}^k-\boldsymbol{z}^{k+1}\right)-\boldsymbol{A}^\top\left(\boldsymbol{A}\boldsymbol{x}^{k-1}-\boldsymbol{z}^k\right)$, [3] where $\boldsymbol{z}^{k+1}$ is chosen from $\operatorname{Arg\,min}_{\boldsymbol{z}\in\Omega}\Psi(\boldsymbol{A}\boldsymbol{x}^k-\boldsymbol{z})$. We choose $\boldsymbol{\eta}^{k+1}\in\partial\mathcal{J}_2(\boldsymbol{x}^k)$, set $\boldsymbol{\zeta}^k:=\boldsymbol{A}^\top\boldsymbol{z}^{k+1}+\boldsymbol{\eta}^{k+1}\in\partial g(\boldsymbol{x}^k)$ [4] and solve subproblems in the following form in each iteration; see [22, Eq. 46]:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n}\left\{\langle\nabla h(\boldsymbol{x}^k)-\boldsymbol{\zeta}^k,\boldsymbol{x}-\boldsymbol{x}^k\rangle+\frac{L_k}{2}\|\boldsymbol{x}-\boldsymbol{x}^k\|^2+P(\boldsymbol{x})\right\}. \tag{38}$$

The above subproblem has a closed-form solution, thanks to $P(\boldsymbol{x})=P_1(\boldsymbol{x})=\lambda\|\boldsymbol{x}\|_1$. Finally, to derive a reasonable termination criterion, we note from the first-order optimality condition of (38) (with $L_k=\bar{L}_k$ determined in [22, Apendix A, Algorithm 2, Step 2]) that

$$-\nabla h(\boldsymbol{x}^{k-1})+\boldsymbol{\zeta}^{k-1}-\bar{L}_{k-1}(\boldsymbol{x}^k-\boldsymbol{x}^{k-1})\in\partial P(\boldsymbol{x}^k).$$

On the other hand, we have (recalling that $f=h$, $P_1=P$ and $P_2=g$) that

$$\partial\Xi(\boldsymbol{x}^k,\boldsymbol{\zeta}^{k-1})=\begin{bmatrix}\nabla h(\boldsymbol{x}^k)+\partial P(\boldsymbol{x}^k)-\boldsymbol{\zeta}^{k-1}\\ -\boldsymbol{x}^k+\partial g^*(\boldsymbol{\zeta}^{k-1})\end{bmatrix}.$$

These together with $\boldsymbol{x}^{k-1}\in\partial g^*(\boldsymbol{\zeta}^{k-1})$ give

$$\begin{bmatrix}\nabla h(\boldsymbol{x}^k)-\nabla h(\boldsymbol{x}^{k-1})-\bar{L}_{k-1}(\boldsymbol{x}^k-\boldsymbol{x}^{k-1})\\ -\boldsymbol{x}^k+\boldsymbol{x}^{k-1}\end{bmatrix}\in\partial\Xi(\boldsymbol{x}^k,\boldsymbol{\zeta}^{k-1}),$$

Thus, we terminate the algorithm when

$$\sqrt{\left(\sqrt{L}\|\boldsymbol{A}(\boldsymbol{x}^k-\boldsymbol{x}^{k-1})\|+\bar{L}_{k-1}\|\boldsymbol{x}^k-\boldsymbol{x}^{k-1}\|\right)^2+\|\boldsymbol{x}^k-\boldsymbol{x}^{k-1}\|^2}<10^{-4}\max\{1,\|\boldsymbol{x}^k\|\},$$

so that we have $\operatorname{dist}(0,\partial\Xi(\boldsymbol{x}^k,\boldsymbol{\zeta}^{k-1}))<10^{-4}\max\{1,\|\boldsymbol{x}^k\|\}$.

[3] Note that $\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^k-\boldsymbol{z}^{k+1})=\nabla h(\boldsymbol{x}^k)-\boldsymbol{\zeta}^k+\boldsymbol{\eta}^{k+1}$ by our choice of $\boldsymbol{\zeta}^k$ in the subproblem (38). Thus, this quantity can be obtained as a by-product when solving (38).

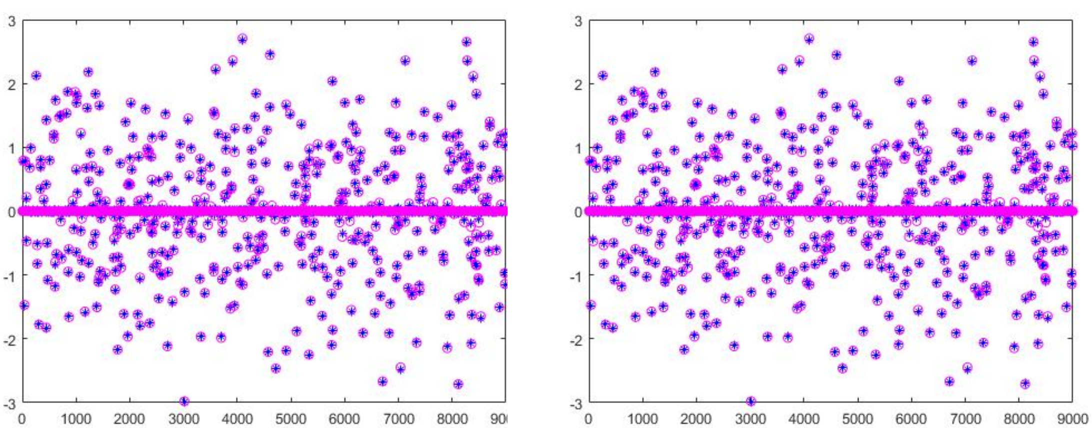
[4] Notice from $\psi_i(s)=\frac{1}{2}(s-b_i)^2$, (33) and the definition of $\boldsymbol{z}^{k+1}$ that $\boldsymbol{A}^\top\boldsymbol{z}^{k+1}\in\partial Q(\boldsymbol{x}^k)$. This together with $\boldsymbol{\eta}^{k+1}\in\partial\mathcal{J}_2(\boldsymbol{x}^k)$ and $g=\mathcal{J}_2+Q$ gives $\boldsymbol{\zeta}^k\in\partial g(\boldsymbol{x}^k)$.

Simulation results: We first generate a matrix $\boldsymbol{A} \in \mathbb{R}^{(m+t)\times n}$ with i.i.d. standard Gaussian entries and then normalize each column of $\boldsymbol{A}$ to have unit norm. Next, we let $\boldsymbol{x}_{\text{true}} \in \mathbb{R}^n$ be an s-sparse vector with s i.i.d. standard Gaussian entries at random positions. Moreover, we choose $\boldsymbol{z} \in \mathbb{R}^{m+t}$ to be the vector with the last t entries being 8 and others being 0. The vector $\boldsymbol{b}$ is then generated as $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}_{\text{true}} - \boldsymbol{z} + \sigma\boldsymbol{\epsilon}$, where $\sigma > 0$ is a noise factor and $\boldsymbol{\epsilon} \in \mathbb{R}^{m+t}$ is a random vector with i.i.d. standard Gaussian entries.

In our numerical test, we consider three different values for λ: 5×10^{-3}, 10^{-3} and 5×10^{-4} in (36). For the same λ value, for each $(m, n, s, t) = (600i, 3000i, 150i, 30i)$, $i = 1, 2, 3$, we generate 20 random instances as described above with $\sigma = 10^{-2}$ and solve the corresponding (36) with $\mu = 0.99$, $p = 0.8s$ and $r \in \{t, 1.1t\}$. Our computational results are reported in Tables 1 and 2. We present the number of iterations (iter), the best function values attained till termination (fval) and CPU times in seconds (CPU), averaged over the 20 random instances. One can see that pDCA$_e$ is always faster than NPG$_{\text{major}}$ and returns slightly smaller function values.

Finally, to illustrate the ability of recovering the original sparse solution by solving (36) with the chosen parameters p, r and λ, we also present in the tables the root-mean-square-deviation (RMSD) $\frac{1}{\sqrt{n}}\|\boldsymbol{x}_{\text{pDCA}_\text{e}} - \boldsymbol{x}_{\text{true}}\|$ for the approximate solution $\boldsymbol{x}_{\text{pDCA}_\text{e}}$ returned by pDCA$_e$ that corresponds to the best attained function value, averaged over the 20 random instances. The relatively small RMSD's obtained suggest that our method is able to recover the original sparse solution approximately. As a further illustration, we also plot $\boldsymbol{x}_{\text{pDCA}_\text{e}}$ (marked by asterisks) against $\boldsymbol{x}_{\text{true}}$ (marked by circles) in Figure 1 below for a randomly generated instance with $m = 1800$, $n = 9000$, $s = 450$ and $t = 90$ (i.e., $i = 3$). We use $\mu = 0.99$, $\lambda = 5 \times 10^{-4}$, $p = 0.8s$, and set $r = t$ and $r = 1.1t$ in Figures 1(a) and 1(b), respectively. One can see that the recovery results are similar even though the r used are different.

Figure 1: Recovery comparison for different r.



(a) Recovery result for $r = t$.

(b) Recovery result for $r = 1.1t$.

Table 1: Numerical results for regression problem with $p = 0.8s$ and $r = t$.

λ	m	n	s	t	RMSD	iter		fval		CPU	
						pDCA$_e$	NPG	pDCA$_e$	NPG	pDCA$_e$	NPG
5e-03	600	3000	150	30	5.0e-03	431	468	3.6365e-02	3.6380e-02	0.7	1.1
	1200	6000	300	60	4.9e-03	422	460	7.1070e-02	7.1093e-02	2.9	4.1
	1800	9000	450	90	5.0e-03	418	439	1.0485e-01	1.0491e-01	6.2	8.5
1e-03	600	3000	150	30	5.4e-03	1276	1966	7.4019e-03	7.4293e-03	2.1	4.6
	1200	6000	300	60	5.5e-03	1254	1956	1.5032e-02	1.5092e-02	8.5	18.1
	1800	9000	450	90	5.5e-03	1298	2013	2.2594e-02	2.2667e-02	18.9	39.7
5e-04	600	3000	150	30	6.0e-03	2361	3844	3.9910e-03	4.0136e-03	3.9	9.0
	1200	6000	300	60	5.8e-03	2367	3890	7.5756e-03	7.6306e-03	16.0	36.1
	1800	9000	450	90	5.8e-03	2311	3841	1.1407e-02	1.1523e-02	33.7	76.5

Table 2: Numerical results for regression problem with $p = 0.8s$ and $r = 1.1t$.

λ	m	n	s	t	RMSD	iter		fval		CPU	
						pDCA$_e$	NPG	pDCA$_e$	NPG	pDCA$_e$	NPG
5e-03	600	3000	150	30	5.1e-03	461	487	3.5891e-02	3.6026e-02	0.8	1.1
	1200	6000	300	60	5.0e-03	454	483	7.0182e-02	7.0291e-02	3.1	4.4
	1800	9000	450	90	5.1e-03	447	469	1.0346e-01	1.0374e-01	6.6	9.1
1e-03	600	3000	150	30	5.4e-03	1530	2067	7.3386e-03	7.3593e-03	2.6	4.9
	1200	6000	300	60	5.6e-03	1485	2053	1.4881e-02	1.4954e-02	10.1	19.1
	1800	9000	450	90	5.5e-03	1550	2114	2.2397e-02	2.2483e-02	22.6	41.8
5e-04	600	3000	150	30	6.0e-03	2837	4114	3.9613e-03	3.9972e-03	4.7	9.7
	1200	6000	300	60	5.8e-03	2874	4061	7.5225e-03	7.5799e-03	19.4	37.7
	1800	9000	450	90	5.9e-03	2778	4027	1.1384e-02	1.1449e-02	40.5	80.1

References

[1] A. Alfons, C. Croux, and S. Gelper. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, 7, 226–248, 2013.

[2] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions invoving analytic features. *Mathematical Programming*, 116, 5–16, 2009.

[3] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35, 438–457, 2010.

[4] H. Attouch, J. Bolte and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Mathematical Programming*, 137, 91–129, 2013.

[5] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27, 265–274, 2009.

[6] J. Bolte, S. Sabach and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146, 459–494, 2014.

[7] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51, 4203–4215, 2005.

[8] E. J. Candès, J. K. Romberg and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59, 1207–1223, 2006.

[9] E. J. Candès, M. B. Wakin and S. P. Boyd. Enhancing spasity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14, 877–905, 2008.

[10] R. E. Carrillo, A. B. Ramirez, G. R. Arce, K. E. Barner and B. M. Sadler. Robust compressive sensing of sparse signals: a review. *EURASIP Journal on Advances in Signal Processing*, 2016.

[11] A. Chambolle and Ch. Dossal. On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization Theory and Applications*, 166, 968–982, 2015.

[12] R. Chartrand. Exact reconstructions of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, 14, 707–710, 2007.

[13] R. Chartrand and W. Yin. Iteratively reweighted algorithms for compressive sensing. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3869–3872, 2008.

[14] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52, 1289–1306, 2006.

[15] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360, 2001.

[16] S. Foucart and M. J. Lai. Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. *Applied and Computational Harmonic Analysis*, 26, 395–407, 2009.

[17] A. Giloni and M. Padberg. Least trimmed squares regression, least median squares regression, and mathematical programming. *Mathematical and Computer Modelling*, 35, 1043–1060, 2002.

[18] P. Gong, C. Zhang, Z. Lu, J. Huang and J. Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. *International Conference on Machine Learning*, 37–45, 2013.

[19] J. Gotoh, A. Takeda and K. Tono. DC formulations and algorithms for sparse optimization problems. *Mathematical Programming*, 169, 141—176, 2018.

[20] J. Hoeting, A. E. Raftery, and D. Madigan. A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics & Data Analysis*, 22, 251–270, 1996.

[21] G. Li and T. K. Pong. Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of Computational Mathematics*, 18, 1199–1232, 2018.

[22] T. Liu, T. K. Pong and A. Takeda. A successive difference-of-convex approximation method for a class of nonconvex nonsmooth optimization problems. To appear in *Mathematical Programming.* DOI:10.1007/s10107-018-1327-8.

[23] P.-L. Loh. Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics*, 45, 866–896, 2017.

[24] Z. Lu and Y. Zhang. Sparse approximation via penalty decomposition methods. *SIAM Journal on Optimization*, 23, 2448–2478, 2013.

[25] R. S. Menjoge and R. E. Welsch. A diagnostic method for simultaneous feature selection and outlier identification in linear regression. *Computational Statistics & Data Analysis*, 54, 3181–3193, 2010.

[26] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24, 227–234, 1995.

[27] D. T. Pham and H. A. Le Thi. Convex analysis approach to DC programming: theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22, 289–355, 1997.

[28] D. T. Pham and H. A. Le Thi. A DC optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization*, 8, 476–505, 1998.

[29] L. F. Polania, R. E. Carrillo, M. Blanco-Velasco and K. E. Barner. Compressive sensing for ECG signals in the presence of electromyography noise. *Proceedings of the 38th Annual Northeast Bioengineering Conference*, 295–296, 2012.

[30] R. T. Rockafellar and R. J-B. Wets. *Variational Analysis*. Springer 1998.

[31] P. J. Rousseeuw. Regression techniques with high breakdown point. *The Institute of Mathematical Statistics Bulletin*, 12, 155, 1983.

[32] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*. John Wiley & Sons, 1987.

[33] R. Saab, R. Chartrand and O. Yilmaz. Stable sparse approximations via nonconvex optimization. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 3885–3888, 2008.

[34] Y. She and A. B. Owen. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106, 626–639, 2011.

[35] E. Smucler and V. J. Yohai. Robust and sparse estimators for linear regression models. *Computational Statistics & Data Analysis*, 111, 116–130, 2017.

[36] R. Tibshirani and J. Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39, 1335–1371, 2011.

[37] H. Tuy. *Convex Analysis and Global Optimization*. Springer, 2016.

[38] Y. Wang, Z. Luo and X. Zhang. New improved penalty methods for sparse reconstruction based on difference of two norms. Available at researchgate. DOI: 10.13140/RG.2.1.3256.3369.

[39] B. Wen, X. Chen and T. K. Pong. Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. *SIAM Journal on Optimization*, 27, 124–145, 2017.

[40] B. Wen, X. Chen and T. K. Pong. A proximal difference-of-convex algorithm with extrapolation. *Computational Optimization and Applications*, 69, 297–324, 2018.

[41] P. Yin, Y. Lou, Q. He and J. Xin. Minimization of ℓ_{1-2} for compressed sensing. *SIAM Journal on Scientific Computing*, 37, A536–A563, 2015.

[42] C. H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38, 894–942, 2010.